BMC Medical Research Methodology

TECHNICAL ADVANCE Open Access

A novel method for controlling unobserved confounding using double confounders

Lu Liu[1,2†], Lei Hou[1,2†], Yuanyuan Yu[1,2], Xinhui Liu[1,2], Xiaoru Sun[1,2], Fan Yang[1,2], Qing Wang[1,2], Ming Jing[1,2], Yeping Xu[3], Hongkai Li[1,2*] and Fuzhong Xue[1,2*]

Abstract

Background: Controlling unobserved confounding still remains a great challenge in observational studies, and a series of strict assumptions of the existing methods usually may be violated in practice. Therefore, it is urgent to put forward a novel method.

Methods: We are interested in the causal effect of an exposure on the outcome, which is always confounded by unobserved confounding. We show that, the causal effect of an exposure on a continuous or categorical outcome is nonparametrically identified through only two independent or correlated available confounders satisfying a non-linear condition on the exposure. Asymptotic theory and variance estimators are developed for each case. We also discuss an extension for more than two binary confounders.

Results: The simulations show better estimation performance by our approach in contrast to the traditional regression approach adjusting for observed confounders. A real application is separately applied to assess the effects of Body Mass Index (BMI) on Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Fasting Blood Glucose (FBG), Triglyceride (TG), Total Cholesterol (TC), High Density Lipoprotein (HDL) and Low Density Lipoprotein (LDL) with individuals in Shandong Province, China. Our results suggest that SBP increased 1.60 (95% CI: 0.99–2.93) mmol/L with per 1- kg/m^2 higher BMI and DBP increased 0.37 (95% CI: 0.03–0.76) mmol/L with per 1- kg/m^2 higher BMI. Moreover, 1- kg/m^2 increase in BMI was causally associated with a 1.61 (95% CI: 0.96–2.97) mmol/L increase in TC, a 1.66 (95% CI: 0.91–55.30) mmol/L increase in TG and a 2.01 (95% CI: 1.09–4.31) mmol/L increase in LDL. However, BMI was not causally associated with HDL with effect value − 0.20 (95% CI: − 1.71–1.44). And, the effect value of FBG per 1- kg/m^2 higher BMI was 0.56 (95% CI: − 0.24–2.18).

Conclusions: We propose a novel method to control unobserved confounders through double binary confounders satisfying a non-linear condition on the exposure which is easy to access.

Keywords: Unobserved confounders, Generalized moment estimate model, Identification, Causal effect

* Correspondence: lihongkaiyouxiang@163.com; xuefzh@sdu.edu.cn
†Lu Liu and Lei Hou are co-first author
[1] Institute for Medical Dataology, Shandong University, 250012 Jinan, Shandong, People's Republic of China
Full list of author information is available at the end of the article

Background

Controlling unobserved confounding is a great challenge when estimating the causal effect of an exposure on an outcome of interest in observational studies [1–4]. Several techniques such as traditional regression model, marginal structure model, adjustment, stratification, inverse probability weighing (IPW), matching based on propensity score cannot deal with unobserved confounding [5–9]. The observed data distribution may have compatibility with many contradictory causal explanations due to the existence of unobserved confounding. In this circumstance, we say the causal estimand is not identified. On the contrary, when the causal estimand can be obtained entirely from observable probability distributions, we say the query is identified.

Some methods have been developed to alleviate the problems caused by unobserved confounding. Instrumental variable analysis (IVA) is the commonly used method to eliminate unobserved confounding [10]. But in practice, choosing valid instruments (IVs) is a stumbling block in IVA [11–13]. The difference-in-differences (DID) is contingent on the availability of repeated outcomes in both periods, but invokes strict parallel trend assumptions, i.e., confounders varying across the groups are time invariant and time-varying confounders are group invariant [14, 15]. Regression discontinuity design (RDD) is a quasi-experimental pretest-posttest design for controlling unobserved confounding by assigning a cutoff or threshold above or below to a treatment [16]. Nevertheless, RDD requires that treatment assignment is sufficiently randomized at the threshold [17]. Negative controls are widely used in epidemiologic practice to detect the presence of unobserved confounding. While a valid negative control outcome needs to be influenced by the same unobserved confounders of the exposure effects on the outcome in view, although not directly influenced by the exposure. But this approach fails to obtain causal estimation of the exposure on the outcome [18]. Moreover, strict assumptions of the methods above usually may be violated in practice and impose restrictions on their generalization.

In this article, we propose a novel method to control unobserved confounding through double confounders with two values satisfying a non-linear condition on the exposure. Under the assumption of ignorable treatment assignment, causal effects can be identified and estimated using commonly generalized moment estimate model. Furthermore, we relax the assumption that observed and unobserved confounders are independent in sensitivity analysis and observe that even when the correlations between binary observed confounders and unobserved confounders are relatively weak, we still obtain the almost unbiased causal effect estimation. Additionally, we explore the statistical properties of this method by a simulation study and compare with the traditional regression approach only adjusting for observed confounders. Finally, we apply this method to a cohort from a follow-up survey (136,895 individuals) from 2007 to 2015 in Jining, China to exam the causal associations of BMI on other factors, including SBP, DBP, FBG, TG, TC, HDL and LDL.

Methods

Notation and preliminaries

Throughout, we let X, Y, C_1, C_2 and U denote the treatment, outcome, two observed confounders and unobserved confounders, respectively (Fig. 1). Following the convention in causal inference, we use $Y(x)$ to denote the potential outcome of Y under an intervention which sets X to x, and the observed outcome Y is a realization of the potential outcome under the exposure actually received: $Y = Y(x)$ when $X = x$. We focus on the average causal effect (ACE) of X on Y which is the difference in expectation of potential outcome at two different exposure levels 0 and 1, for instance, $ACE_{X\rightarrow Y} = E(Y(1) - Y(0))$ for a binary exposure. The conditional ignorability assumption $Y(x) \perp x \mid C_1, C_2$ is conventionally made in causal inference, but it does not hold in the present of unmeasured confounding U. In this case, latent ignorability $Y(x) \perp x \mid C_1, C_2, U$ is more reasonable, allowing for an unobserved confounder U that captures the source of non-ignorability of the exposure mechanism. The model with continuous exposure and outcome can be developed as follows.

$$X = F(C_1, C_2) + \varphi(U, \varepsilon_X) \tag{1}$$

$$Y = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \phi(U, \varepsilon_Y) \tag{2}$$

where ε_Y and ε_X are two mutually independent random errors with means 0 and variances σ_Y^2 and σ_X^2 respectively, $\varphi(\cdot)$ and $\phi(\cdot)$ are two arbitrary functions. The causal effect of X on Y is interpreted as β_1. Similarly, for binary exposure and outcome, we can also construct corresponding linear probability model (LPM) for $P(X = 1 | C_1, C_2, U)$, $P(Y = 1 | C_1, C_2, U)$, respectively. In addition, the "complementary log link" of transformation of Y (i.e. $-\log(1 - Risk)$, $Risk$ denotes cumulative completion risk) is also appropriate for our method [19].

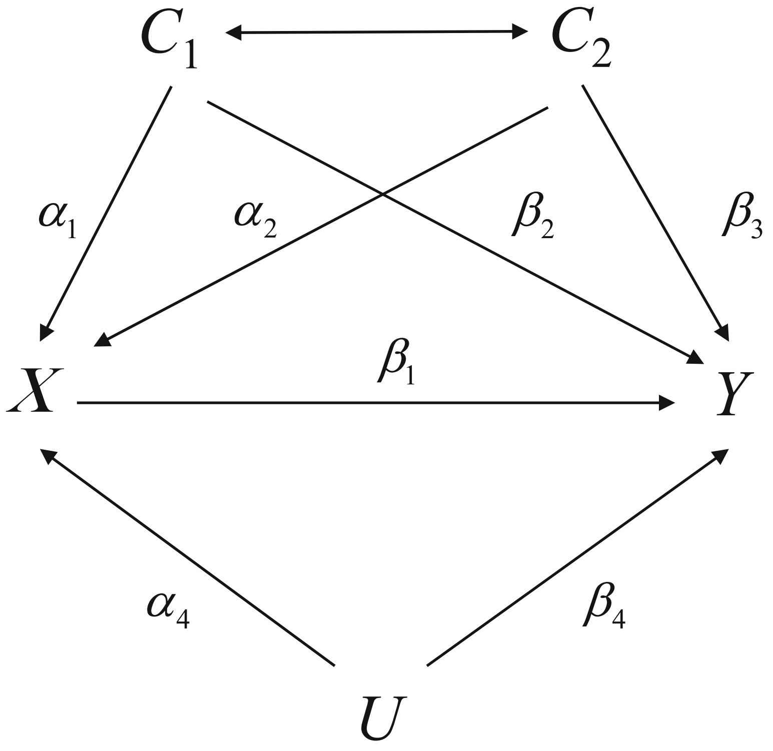

Fig. 1 Causal diagram for X (treatment) and Y (outcome) with C_1, C_2 (observed confounders) and U (other unobserved confounders)

Identification and estimation of causal effects with double binary confounders

Obviously, the estimation of regression model adjusting for C_1 and C_2 is biased when there exists an unobserved confounder U. Unfortunately, the no unobserved confounding assumption usually not be satisfied in practical studies. Next, we explain how to identify and estimate the causal effect by relaxing the no unobserved confounding assumption in four cases with different types of exposure and outcome (continuous or binary). It will be shown below that causal effect β_1 in (2) is not identifiable if the function $F(X| C_1, C_2)$ is linear with respect to C_1 and C_2.

Within the causal framework provided by Fig. 1, we propose four assumptions and discuss the necessary and sufficient condition for identification of parameters in the model (2).

Assumption 1: $E(\phi(U, \varepsilon_Y)) = 0$.

Assumption 2: $(C_1, C_2) \perp \phi(U, \varepsilon_Y)$, i.e. (C_1, C_2) are not associated with any confounder (U) of the exposure–outcome relationship and random errors ε_Y.

Assumption 3: The effect of (X, C_1, C_2) on Y is linear (or no interaction).

Assumption 4: The effect of (C_1, C_2) on X is non-linear (e.g. with an interaction effect or the quadratic term of C_1 and C_2 on X).

Under Assumptions 1 and 2, $E(\phi(U, \varepsilon_Y)| C_1, C_2) = 0$ is satisfied. The **Assumption 2** is the same as the exchangeability assumption in instrumental variable (IV) analysis – an IV is not associated with any confounder of the exposure–outcome relationship. In addition, a valid IV must have a main effect on X and no direct effect on Y, which are called the Relevance and the Exclusion restriction assumption, respectively [20]. In our method, (C_1, C_2) can be regarded as two near-IVs with a non-linear effect on X and a linear effect on Y as stated in the **Assumption 3** and **4**.

Theorem 1: *The causal effect β_1 of X on Y in model (2) is identifiable if and only if the Assumptions 1–4 are satisfied.*

See Appendix A for the proof of Theorem 1.

Causal effect is identified based on above four assumptions. And various estimation approaches have been applied to estimate the causal effect, including moment estimation, maximum likelihood estimation and generalized moment estimate model (GMM) with different assumptions [21–23]. In this section, we aim to find an efficient estimation of causal effect in the model (2).

In our estimation approach, the pivotal “orthogonality condition” is the independence $(C_1, C_2) \perp \phi(U, \varepsilon_Y)$, which implies the following equation:

$$E\left[\left(Y-\beta_0-\beta_1X-\beta_2C_1-\beta_3C_2\right)f(C_1,C_2)\right]=\mathbf{0}, \quad (3)$$

where $f(\cdot) = (f_1(\cdot), \cdots, f_K(\cdot))^T$ is an arbitrary vector function and $\mathbf{0}$ is a $K \times 1$ zero vector.

For the case of confounders C_1 and C_2 with two values, we chose the function $f^*(C_1, C_2) = (\delta(C_1 = 0, C_2 = 0), \delta(C_1 = 0, C_2 = 1), \delta(C_1 = 1, C_2 = 0), \delta(C_1 = 1, C_2 = 1))^T$, we have

$$E\left[\left(Y-\beta_0-\beta_1X-\beta_2C_1-\beta_3C_2\right)f^*(C_1,C_2)\right]=\mathbf{0}, \quad (4)$$

where $\mathbf{0}$ is a 4×1 zero vector. Define $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$,

$$R^* = \begin{pmatrix} E(Y\delta(C_1=0,C_2=0)) \\ E(Y\delta(C_1=0,C_2=1)) \\ E(Y\delta(C_1=1,C_2=0)) \\ E(Y\delta(C_1=1,C_2=1)) \end{pmatrix},$$

$$Q^* = \begin{pmatrix} E(\delta(C_1=0,C_2=0)) & E(X\delta(C_1=0,C_2=0)) & E(C_1\delta(C_1=0,C_2=0)) & E(C_2\delta(C_1=0,C_2=0)) \\ E(\delta(C_1=0,C_2=0)) & E(X\delta(C_1=0,C_2=1)) & E(C_1\delta(C_1=0,C_2=1)) & E(C_2\delta(C_1=0,C_2=1)) \\ E(\delta(C_1=0,C_2=0)) & E(X\delta(C_1=1,C_2=0)) & E(C_1\delta(C_1=1,C_2=0)) & E(C_2\delta(C_1=1,C_2=0)) \\ E(\delta(C_1=0,C_2=0)) & E(X\delta(C_1=1,C_2=1)) & E(C_1\delta(C_1=1,C_2=1)) & E(C_2\delta(C_1=1,C_2=1)) \end{pmatrix},$$

the Eq. (4) can be rewritten as $R^* - Q^*\boldsymbol{\beta} = 0$. Solve for GMM estimator

$$\widehat{\boldsymbol{\beta}^*} = \widehat{Q^*}^{-1}\widehat{R^*},$$

where the elements of $\widehat{Q^*}$ and $\widehat{R^*}$ are sample means about X, C_1, C_2 and C_1, C_2, Y, respectively. Thus $\widehat{\boldsymbol{\beta}^*}$ is a valid estimator only if Q^* has full rank. The determinant of the matrix Q^* is

$$\begin{aligned} \det(Q^*) = {} & P(C_1=0,C_2=0)P(C_1=0,C_2=1) \\ & P(C_1=1,C_2=0)P(C_1=1,C_2=1) \\ & \cdot\{[E(X|C_1=0,C_2=1)-E(X|C_1=0,C_2=0)]- \\ & [E(X|C_1=1,C_2=1)-E(X|C_1=1,C_2=0)]\}. \end{aligned}$$

Note that the nonlinearity of $E[X| C_1, C_2]$ with respect to C_1 and C_2 implies

$$\begin{aligned} & [E(X|C_1=0,C_2=1)-E(X|C_1=0,C_2=0)]- \\ & [E(X|C_1=1,C_2=1)-E(X|C_1=1,C_2=0)]\neq 0. \end{aligned}$$

Furthermore, the estimated causal effect $\widehat{\beta_1}$ is

$$\widehat{\beta_1} = \frac{[E(Y|C_1=1,C_2=1)-E(Y|C_1=1,C_2=0)]-[E(Y|C_1=0,C_2=1)-E(Y|C_1=0,C_2=0)]}{[E(X|C_1=0,C_2=1)-E(X|C_1=0,C_2=0)]-[E(X|C_1=1,C_2=1)-E(X|C_1=1,C_2=0)]}.$$

The estimated causal effect $\widehat{\beta_1}$ is reasonable compared with the "ratio estimator" of a binary IV when C_1 and C_2 are regarded as two near-IVs. Denote the sub-sample average of y and x by $\overline{y}_1$ and $\overline{x}_1$ when $z = 1$ and by $\overline{y}_0$ and $\overline{x}_0$ when $z = 0$. Then $\Delta y/\Delta z = \overline{y}_1 - \overline{y}_0$ and $\Delta x/\Delta z = \overline{x}_1 - \overline{x}_0$, and "ratio estimator" of an IV [24] is

$$\hat{\beta}_{IV} = \frac{\overline{y}_1 - \overline{y}_0}{\overline{x}_1 - \overline{x}_0}.$$

Define $g(\boldsymbol{\beta}) = E[(Y - \beta_0 - \beta_1X - \beta_2C_1 - \beta_3C_2)f(C_1, C_2)]$, and $\boldsymbol{\beta}$ can be identified only when $rank(Q) = l + 2$, $l = 2$, where l denotes the number of the observed confounders. For any $f(\cdot)$ that makes $rank(Q) = l + 2$, a GMM estimate of $\boldsymbol{\beta}$ is

$$\begin{aligned} \hat{\boldsymbol{\beta}} = {} & \arg\ \min\left\{\hat{g}(\boldsymbol{\beta})'\hat{W}\hat{g}(\boldsymbol{\beta})\right\} \\ = {} & \left(\hat{Q}'\hat{W}\hat{Q}\right)^{-1}\hat{Q}'\hat{W}\hat{R}, \end{aligned} \quad (5)$$

where $\hat{W}$ is denoted as a symmetric and positive definite weight matrix for $K \geq (l + 2)$, $K = \text{rank}(Q)$. The elements of $\hat{g}(\boldsymbol{\beta}), \hat{Q}$ and $\hat{R}$ are sample means of the corresponding elements of $g(\boldsymbol{\beta})$, Q and R respectively [20].

If $\hat{W} \rightarrow W$ as $N \rightarrow \infty$ in probability where W is positive semi-definite and N denotes the sample size. The main general properties about GMM estimators under the appropriate regularity conditions are that,

1. $\hat{\boldsymbol{\beta}} \rightarrow \boldsymbol{\beta}$ in probability as $N \rightarrow \infty$ where $\boldsymbol{\beta}$ denotes the true parameter.
2. $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a normal distribution with mean zero and variance $(Q^TWQ)^{-1}Q^TWE[f(C_1, C_2)f(C_1, C_2)^T]WQ(Q^TWQ)^{-1}\sigma^2_{res}$, where σ^2_{res} is the variance of $\phi(U, \varepsilon_Y)$.

Then we describe a GMM estimator with the efficient instrument (a function of C_1, C_2) proposed by Newey and McFadden [25], which has the minimum variance among all estimators satisfying the Eq. (4). It also shown that there is an efficient estimator $\hat{\boldsymbol{\beta}}$ when the instrument $f(C_1, C_2)$ is defined as

$$\begin{aligned} f^{\text{eff}}(C_1,C_2) = {} & \left\{E\left[\frac{\partial\psi(X,C_1,C_2,Y)}{\partial\boldsymbol{\beta}}\right]|C_1,C_2\right\}^T \\ = {} & (1, E[X|C_1,C_2], C_1, C_2)^T. \end{aligned}$$

From (5), the efficient GMM estimator of $\boldsymbol{\beta}$ is defined as $\hat{\boldsymbol{\beta}}^{\text{eff}} = ((\hat{Q}^{\text{eff}})^T\hat{W}\hat{Q}^{\text{eff}})^{-1}(\hat{Q}^{\text{eff}})^T\hat{W}\hat{R}^{\text{eff}}$, where $\hat{Q}^{\text{eff}}$ and $\hat{R}^{\text{eff}}$ are the sample means of

$$Q^{\text{eff}} = \begin{bmatrix} 1 & E(X) & E(C_1) & E(C_2) \\ E(X) & E\left[E(X|C_1,C_2)^2\right] & E(C_1X) & E(C_2X) \\ E(C_1) & E(C_1X) & E(C_1^2) & E(C_1C_2) \\ E(C_2) & E(C_2X) & E(C_1C_2) & E(C_2^2) \end{bmatrix}$$

$$R^{\text{eff}} = \begin{bmatrix} E(Y) \\ E[YE(X|C_1,C_2)] \\ E(C_1Y) \\ E(C_2Y) \end{bmatrix}.$$

The matrix Q^{eff} equals $E[f(C_1, C_2)f(C_1, C_2)^T]$ proved in Appendix B and has full rank if the nonlinearity condition in Theorem 1 holds. Since Q^{eff} and the positive definite $\hat{W}$ for $l = 2$ are full rank, $\hat{\boldsymbol{\beta}}^{\text{eff}}$ is a valid estimator which can be simplified as

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{\text{eff}} &= \left(\hat{Q}^{\text{eff}}\right)^{-1}\hat{W}^{-1}\left[\left(\hat{Q}^{\text{eff}}\right)^{T}\right]^{-1}\left(\hat{Q}^{\text{eff}}\right)^{T}\hat{W}\hat{R}^{\text{eff}} \\ &= \left(\hat{Q}^{\text{eff}}\right)^{-1}\hat{R}^{\text{eff}}.\end{aligned} \tag{6}$$

We find $\hat{\boldsymbol{\beta}}^{\text{eff}}$ does not rely on the choice of $\hat{W}$ at this point. From the above property of the GMM estimator, the asymptotic variance of $\hat{\boldsymbol{\beta}}^{\text{eff}}$ is easily obtained by Q^{eff} and σ^2_{res}

$$\begin{aligned}AVAR\left(\hat{\boldsymbol{\beta}}^{\text{eff}}\right) = \left(\hat{Q}^{\text{eff}}\right)^{-1}E\left[f^{\text{eff}}(C_1,C_2)f^{\text{eff}}(C_1,C_2)^{T}\right] \\ \left(\hat{Q}^{\text{eff}}\right)^{-1}\sigma^2_Y = \left(\hat{Q}^{\text{eff}}\right)^{-1}\sigma^2_{res}.\end{aligned}$$

Additionally, we show for two binary observed confounders that any $f(\cdot)$ which makes the Eq. (4) have the unique solution leads to the same estimator of parameters as that obtained by $f^{*}(C_1, C_2)$ in Appendix B. Therefore, for the case of two binary observed confounders, our estimator $\widehat{\boldsymbol{\beta}}^{*}$ is efficient and it is not necessary to choose an extra function $f(\cdot)$ to improve the efficiency.

Extend to the case of more than two binary confounders

Note that the method proposed is an easy tool to detect the causal effect in the absence of enough confounding information. Because of a growing appreciation of the power gains of multivariate association analyses, more than two covariates are generally selected to analysis in practice. In this section, for the case of more than two binary confounders, we discuss the identification of parameter β_1 in the following model

$$\begin{aligned}Y = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \cdots + \beta_{l+1} C_l \\ + \phi(U, \varepsilon_Y).\end{aligned} \tag{7}$$

Define $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_{l+1})^T$, Eq. (7) can be rewritten as $Q \times \boldsymbol{\beta} = R$, where $Q = (1, E(X|\, C_1, C_2, \cdots C_l), C_1, C_2, \cdots C_l)$ denotes a $2^l \times (l+2)$ matrix, $\boldsymbol{\beta}$ is a $(l+2)$ vector. According to the equation $Q \times \boldsymbol{\beta} = R$, parameter β_1 is identifiable if $K = \text{rank}(Q) \geq (l+2)$, such that the matrix Q has full column rank. We extend **Assumptions 2–4** to the case of more than two binary confounders as follows:

Assumption 2*: $(C_1, C_2, \cdots, C_l) \perp \phi(U, \varepsilon_Y)$, i.e. $(C_1, C_2, \cdots, C_l)$ are not associated with any confounder (U) of the exposure–outcome relationship and random errors ε_Y.

Assumption 3*: The effect of $(X, C_1, C_2, \cdots, C_l)$ on Y is linear.

Assumption 4*: The effect of $(C_1, C_2, \cdots, C_l)$ on X is non-linear (e.g. with an interaction effect or the quadratic term of $(C_1, C_2, \cdots, C_l)$ on X).

Thus, under the **Assumption 1** and the **Assumptions 2*-4***, the causal effect of X on Y can be identified.

Similar to the case of two binary observed confounders C_1 and C_2, different $f(\cdot)$ for C_1, C_2, $\cdots$, C_l in (7) leads to different estimators. Here, we have

$$E\left[(Y - \beta_0 - \beta_1 X - \beta_2 C_1 - \beta_3 C_2 - \cdots - \beta_{l+1} C_l) f(C_1, C_2, \cdots, C_l)\right] = \mathbf{0}. \tag{8}$$

Under the conditional expectation (8), the efficient GMM estimator $\hat{\boldsymbol{\beta}}^{\text{eff}}$ of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}^{\text{eff}} = (\hat{Q}^{\text{eff}})^{-1}\hat{R}^{\text{eff}}$, with

$$\begin{aligned}f^{\text{eff}}(C_1, \cdots C_l) &= \left\{E\left[\frac{\partial\psi(X, C_1, \cdots C_l, Y)}{\partial\boldsymbol{\beta}}\right]|C_1, \cdots C_l\right\}^T \\ &= (1, E[X|C_1, \cdots C_l], C_1, \cdots C_l)^T.\end{aligned}$$

From the above property of the GMM estimator, we can obtain the asymptotic variance of $\hat{\boldsymbol{\beta}}^{\text{eff}}$ in this case

$$\begin{aligned}AVAR\left(\hat{\boldsymbol{\beta}}^{\text{eff}}\right) = \left(\hat{Q}^{\text{eff}}\right)^{-1}E\left[f^{\text{eff}}(C_1, \cdots, C_l)f^{\text{eff}}(C_1, \cdots, C_l)^{T}\right] \\ \left(\hat{Q}^{\text{eff}}\right)^{-1}\sigma^2_Y = \left(\hat{Q}^{\text{eff}}\right)^{-1}\sigma^2_{res},\end{aligned}$$

where $\hat{Q}^{\text{eff}}$ and $\hat{R}^{\text{eff}}$ are the sample means of

$$Q^{eff} = \begin{bmatrix} 1 & E(X) & E(C_1) & \cdots & E(C_l) \\ E(X) & E\left[E(X|C_1, \cdots, C_l)^2\right] & E(C_1X) & \cdots & E(C_lX) \\ E(C_1) & E(C_1X) & E\left(C_1^2\right) & \cdots & E(C_1C_l) \\ \vdots & \vdots & \vdots & & \vdots \\ E(C_l) & E(C_lX) & E(C_1C_l) & \cdots & E\left(C_l^2\right) \end{bmatrix},$$

$$R^{eff} = \begin{bmatrix} E(Y) \\ E[YE(X|C_1, \cdots, C_l)] \\ E(C_1Y) \\ \vdots \\ E(C_lY) \end{bmatrix}.$$

Simulations

In order to investigate the performance of our method in simulation study, as well as to determine in which scenario it performs well or badly in comparison with the traditional regression approach adjusting for observed confounders, we perform simulations with four cases. In our simulations, data are generated based on the causal diagram depicted in Fig. 1.

In each case (exposure and outcome are continuous or binary respectively, denoted as Simulation A: exposure and outcome are both continuous; Simulation B: exposure is binary while outcome is continuous; Simulation C: exposure is continuous while outcome is binary; Simulation D: exposure and outcome are both binary), we take into account two scenarios where C_1, C_2 and U are independent, C_1, C_2 and U are correlated respectively (For Simulation A, two scenarios denotes as Simulation A1, Simulation A2, respectively, similar notation for Simulation B, C, D). We performed simulation to compare performances of three methods including crude association without adjusting for C_1, C_2 and U (model 1), model of adjusting for C_1 and C_2 (model 2) and the method we proposed here (model 3). We simulated baseline covariates and a quantitative exposure in a large population consisting of 2000 subjects in Simulation A, 10000 subjects in Simulation B and 100,000 subjects in Simulation C, D.

The data of Simulation A1 were generated for each individual in the following procedure:

1. $C_1 \sim Bernoulli(N, 0.5)$, $C_2 \sim Bernoulli(N, 0.5)$.
2. $U \sim N(0, 1)$.
3. $\varepsilon_X \sim U(0, 1)$, $\varepsilon_Y \sim U(0, 1)$.
4. $X = 0.6C_1 + 0.4C_2 + 0.6C_1C_2 + 0.2U + \varepsilon_X$.
5. $Y = X + 0.4C_1 + 0.3C_2 + 0.1U + \varepsilon_Y$.

The data of Simulation A2 were generated for each individual in the following procedure:

1. $(C_1, C_2, U) \sim N(\mu, \Sigma), \mu = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma = \begin{bmatrix} 1 & d & c \\ d & 1 & c \\ c & c & 1 \end{bmatrix}, c = 0.1, d = 0.2.$
2. Draw C_1 from the first column of multivariate normal distribution, and discretize it into a binary variable with probability$P(C_1 = 1) = 0.5$.
3. Draw C_2 from the second column of multivariate normal distribution, and discretize it into a binary variable with probability$P(C_2 = 1) = 0.5$.
4. Draw U from the third column of multivariate normal distribution.
5. $\varepsilon_X \sim U(0, 1)$, $\varepsilon_Y \sim U(0, 1)$.
6. $X = 0.6C_1 + 0.4C_2 + 0.6C_1C_2 + 0.2U + \varepsilon_X$.
7. $Y = X + 0.4C_1 + 0.3C_2 + 0.1U + \varepsilon_Y$.

The data of Simulation B1-B2, Simulation C1-C2, Simulation D1-D2 were similarly generated for each individual. Details are listed in the Appendix C.

Data application

Numerous epidemiological studies have evaluated the relationships between BMI and SBP, DBP, FBG, TG, TC, HDL and LDL, respectively, but the causal effects are inconclusive due to the existence of unobserved confounding [26–28]. Hence, we use the method proposed in this article to evaluate the potential causal effect of BMI on SBP, DBP, FBG, TG, TC, HDL and LDL compared with the traditional regression approach adjusting for age (discrete) and gender, adjusting for age (continuous) and gender and adjusting for age (discrete), gender and other health factors (including SBP, DBP, FBG, TC, TG, HDL and LDL) using data from a follow-up survey in Jining, Shandong Province.

Results

Simulation results

We separately varied across the main effect value α_1 of C_1 on X, the main effect value α_2 of C_2 on X, the interaction effect α_3 of C_1 and C_2 on X, the causal effect β_1 of X on Y, the confounding effect β_2 of C_1 on Y, the correlation coefficient c between C_1 and C_2 as well as correlation coefficient d between C_1 and U.

We changed one parameter at a time while keeping all others at their basic values, when $N = 2000$ for Simulation A1-A2, $N = 10{,}000$ for Simulation B1-B2, $N = 100{,}000$ for Simulation C1-C2 and $N = 100{,}000$ for Simulation D1-D2. Results showed the estimated bias, standard error (*SE*) and the mean squared error (*MSE*) from the three models for varied effects of C_1 on X, the interaction effects of C_1 and C_2 on X, U on X, X on Y, C_1 on Y, U on Y, and the correlations among C_1, C_2 and U. The results showed the estimates of the model without any adjustment (model 1) and the model adjusting for both C_1 and C_2 (model 2) were biased. When we used the method we proposed (model 3) to analyze simulated data sets, the estimates were unbiased and had acceptable standard error. And our method had better *MSE* than other methods (i.e. model 1 and model 2). Figures 2 and 3 showed the results of Simulation A1, Fig. 4 showed the partial results of Simulation A2, the rest of results of Simulation A2 and other six different scenarios were showed in Figure S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11, S12, S13 (Figure S1-S2 for Simulation A2, Figure S3-S4 for Simulation B1, Figure S5, S6, S7 for Simulation B2, Figure S8-S9 for Simulation C1, Figure S10 for Simulation C2, Figure S11-S12 for Simulation D1, Figure S13 for Simulation D2). Specifically, when varying across the effects of C_1 on Y, the biases of unadjusted model (model 1) had a significant positive linear correlation with β_2 and the ones of adjustment for C_1 and C_2 (model 2) remained stable. When varying across the effects of U on Y, the biases of both model 1 and model 2 linearly increased. Our method still had unbiased causal effect estimates and lower *MSE* than other two methods. Similarly, when varying across the effects of C_1, U on X, respectively, the biases of model 1 and model 2 monotonically varied. Furthermore, the larger interaction effect of C_1 and C_2 on X, the causal effect estimation had higher precision for our proposed method (model 3). In order to make sure model 3 remains the smallest *MSE* among three models, we suggested the interaction effect of C_1 and C_2 on X moderately larger. Certainly, the bias did not change significantly from the basic scenario in any of the models when we changed the causal effect of X on Y. Moreover, when there existed an correlation between observed confounders C_1 and C_2, our method still got unbiased causal effect estimated. We still obtained the almost unbiased

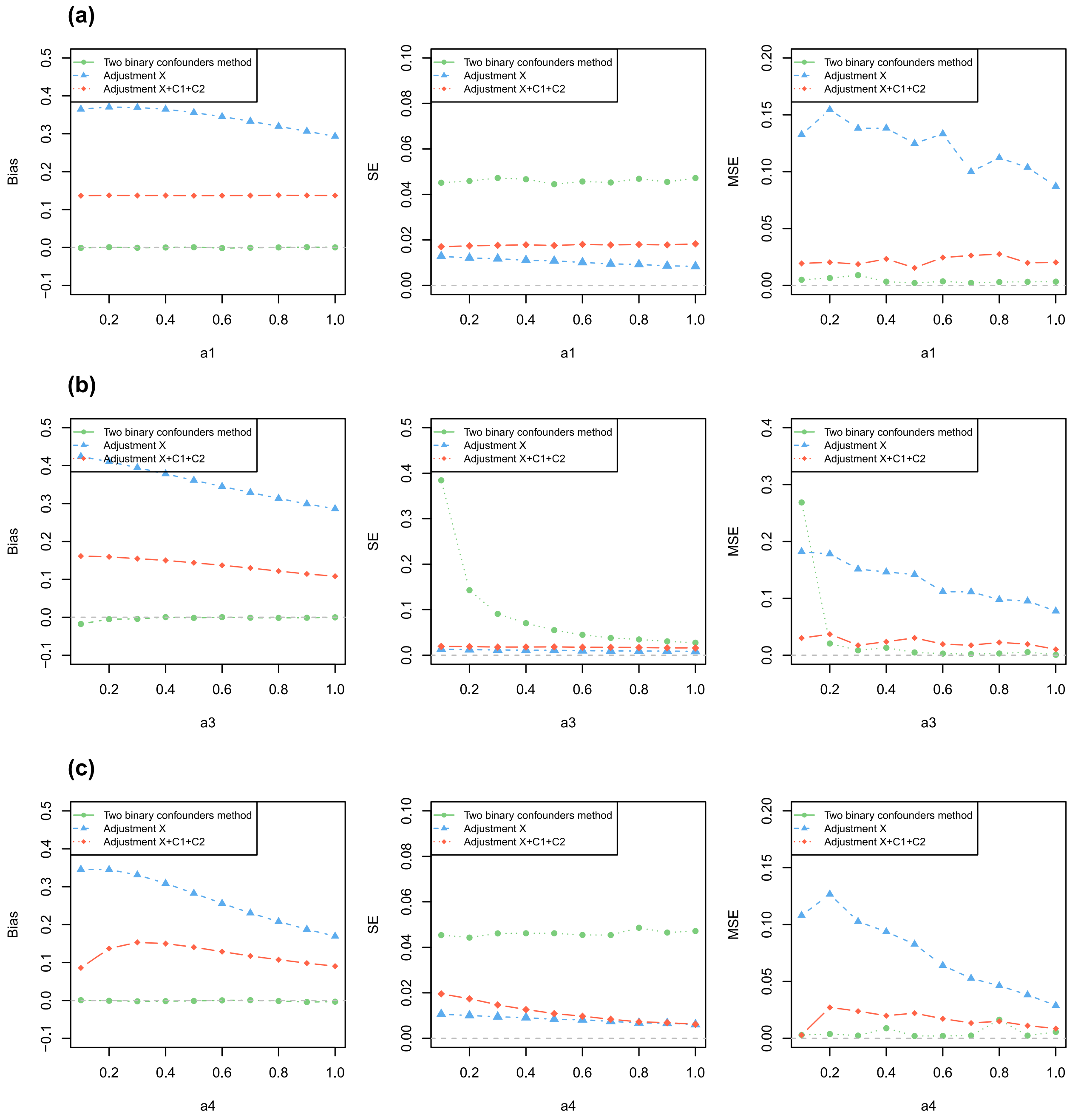

Fig. 2 The Simulation A1 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of **a** C_1 on *X*, **b** the interaction effect of C_1 and C_2 on *X*, **c** *U* on *X*

causal effect estimation if $P(C_1 = 1) = P(C_2 = 1) = 0.5$ when observed confounder C_1 (or C_2) and unobserved confounders *U* were correlated.

Compared to the situation of continuous variables, sample size *N* needs to be larger for discrete exposure and outcome variable. Under the linear probability model setting of X, C_1, C_2, U on Y (i.e. $0 \le P(Y| X, C_1, C_2, U) \le 1$), the effects of C_1 on X, C_2 on X, the interaction effect of C_1 and C_2 on X, U on X should be relatively small. Furthermore, our method had better *MSE* than model 2, and was similar with model 1, the estimates remained unbiased. We use R program (version 3.6.1) to reproduce all simulations and analyses which are available on Github (https://github.com/LULIU1816/Two-binary-confounder).

Data application results

In this section, we used the method proposed in this article to evaluate the potential causal effects of BMI on

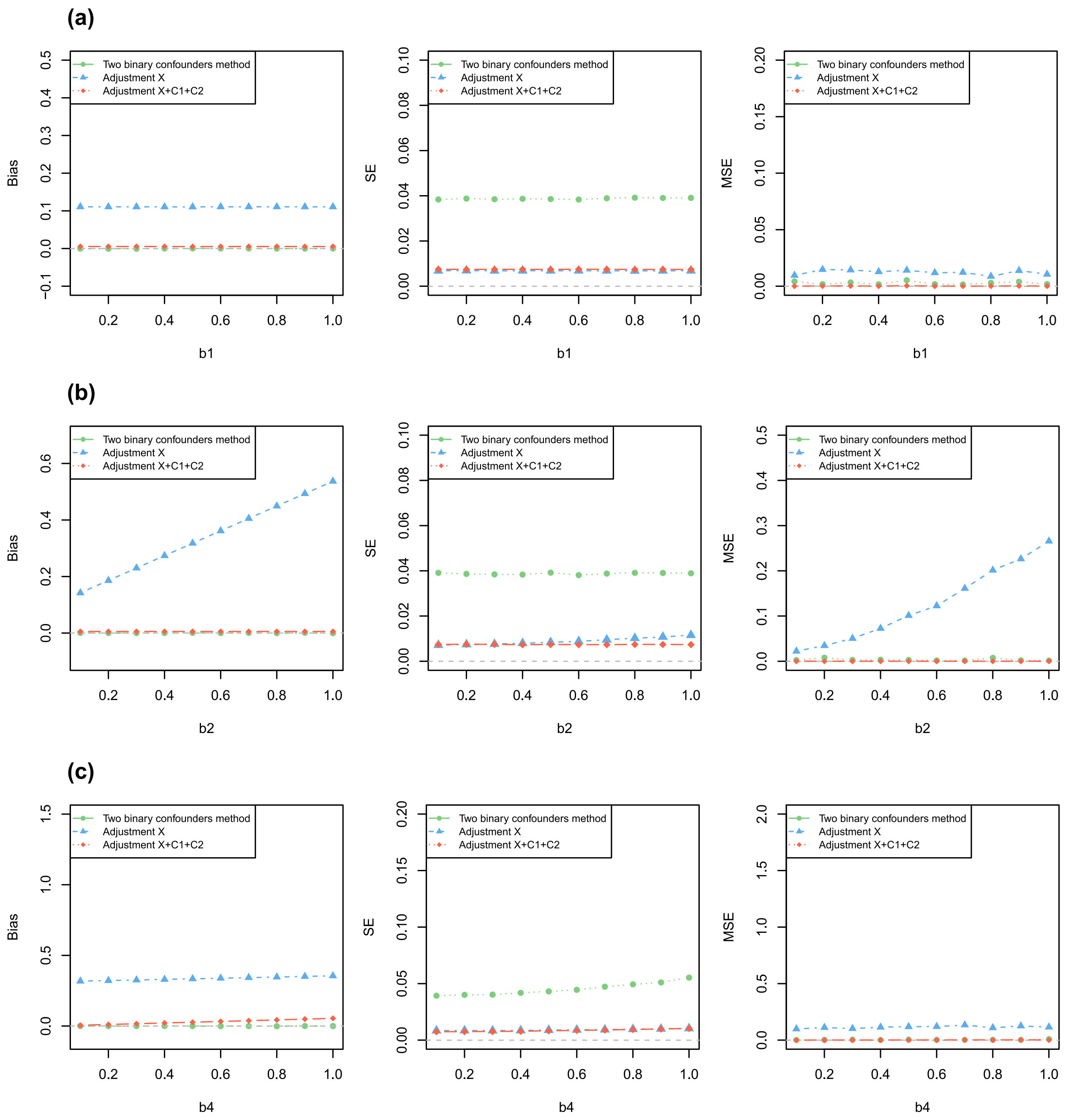

Fig. 3 The Simulation A1 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of **a** *X* on *Y*, **b** C_1 on *Y*, **c** *U* on *Y*

SBP, DBP, FBG, TG, TC, HDL and LDL using data from a follow-up survey in Jining, Shandong Province. Furthermore, we selected age and gender as near-IVs compared with the traditional regression approaches adjusting for age (discrete) and gender, adjusting for age (continuous) and gender and adjusting for age (discrete), gender and other health factors (including SBP, DBP, FBG, TC, TG, HDL and LDL). The cohort recruited 136, 895 individuals aged 20- years between 2007 and 2015. In order to avoid reverse causation, we selected the data of BMI, age, gender in 2014 and SBP, DBP, FBG, TG, TC, HDL and LDL in 2015($N = 7013$), respectively. Age and gender were divided into two discrete variables 0, 1. Age was discretized by median 40 as 0 and 1. Subject characteristics showed at Table 1. Proportion of female was 41.01%. Kolmogorov-Smirnov test is one of normality tests when the sample size over 5000. The median SBP was 120 (95% CI: 109–160) mmHg, the median DBP was 75 (95% CI: 66–101) mmHg. The median FBG was 5.20 (95% CI: 4.80–8.20) mmol/L, the median TC

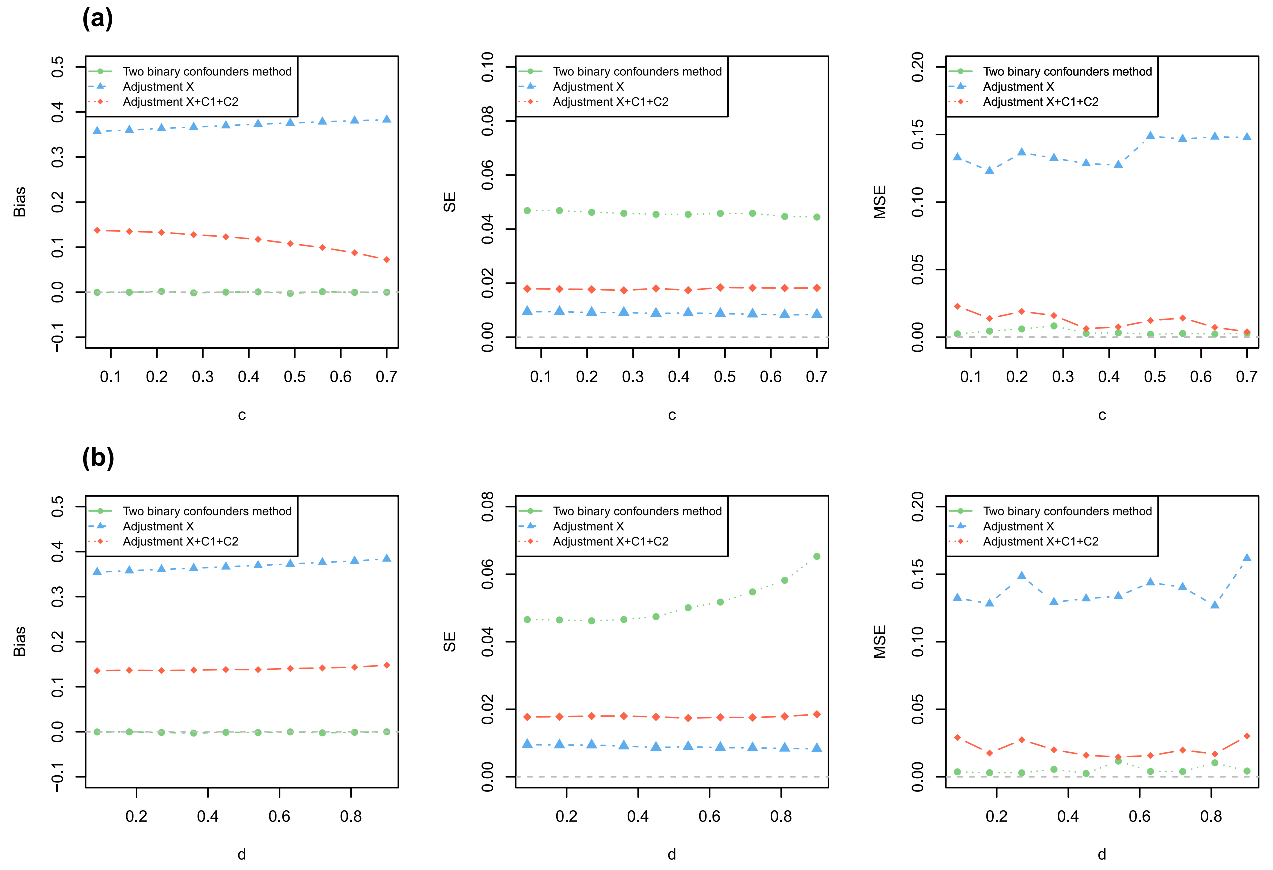

Fig. 4 The Simulation A2 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of **a** the correlation between *X* and C_1, **b** the correlation between C_1 and C_2

was 4.64 (95% CI: 4.01–6.76) mmol/L, the median TG was 1.04 (95% CI: 0.60–4.67) mmol/L, the median HDL was 1.29 (95% CI: 1.10–1.87) mmol/L, the median LDL was 2.76 (95% CI: 2.23–4.30) mmol/L. And, the median BMI was 24.68 (95% CI: 21.83–32.07) kg/m^2.

Table 1 Subjects characteristics in Jining, Shandong Province

Variables	*N* = 7013	P[a]
Age (years, Median(95% CI))	40(31–70)	$< 2.2\times10^{-16}$
Female (n, %)	2876 (41.01%)	
SBP (mmHg, Median(95% CI))	120 (109–160)	$< 2.2\times10^{-16}$
DBP (mmHg, Median(95% CI))	75 (66–101)	$< 2.2\times10^{-16}$
FBG (mmol/L, Median(95% CI))	5.20 (4.80–8.20)	$< 2.2\times10^{-16}$
TC (mmol/L, Median(95% CI))	4.64 (4.01–6.76)	$< 2.2\times10^{-16}$
TG (mmol/L, Median(95% CI))	1.04 (0.60–4.67)	$< 2.2\times10^{-16}$
HDL (mmol/L, Median(95% CI))	1.29 (1.10–1.87)	$< 2.2\times10^{-16}$
LDL (mmol/L, Median(95% CI))	2.74 (2.23–4.30)	$< 2.2\times10^{-16}$
BMI (kg/m^2, Median(95% CI))	24.68 (21.83–32.07)	6.96×10^{-07}

[a]Kolmogorov-Smirnov test

Results of our proposed method and the traditional regression approach adjusting for age and gender are showed in Table 2. Firstly, we examine whether two independent or correlated available confounders satisfy the non-linear condition on the exposure. Results confirmed the interaction between age and gender on BMI ($P = 2.82 \times 10^{-9}$). Traditional regression approach implied the significant associations between SBP, DBP, FBG, TG, TC, HDL, LDL and BMI, respectively. BMI had obviously positive correlation with SBP, DBP, FBG, TG, TC and LDL while obviously negative correlation with HDL. However, our proposed method revealed the causal effect of BMI on indicators including SBP, DBP, TG, TC, LDL. In addition, the causal effects of BMI on HDL and FBG were not significant. SBP increased 1.60 (95% CI: 0.99–2.93) mmol/L with per 1- kg/m^2 higher BMI and DBP increased 0.37 (95% CI: 0.03–0.76) mmol/L with per 1- kg/m^2 higher BMI. Moreover, 1- kg/m^2 increase in BMI had potential causality with a 1.61-SD increase in TC (*β*, 1.61; 95% CI: 0.96–2.97), a 1.66-SD increase in TG (*β*, 1.66; 95% CI: 0.91–55.30) and a 2.01-SD increase in LDL (*β*, 2.01; 95% CI: 1.09–4.31). However, BMI had

Table 2 The causal effect of BMI on SBP, DBP, FBG, TG, TC, HDL and LDL

	double confounding variables model			regression adjusting for age (discrete) and gender		
	β	SE	95%CI	β	SE	*P*-value
BMI → SBP	1.60	0.62	0.99–2.93	0.25	0.01	3.83×10^{-74}
BMI → DBP	0.37	0.18	0.03–0.76	0.22	0.01	3.90×10^{-51}
BMI → FBG	0.56	0.86	−0.24-2.18	0.11	0.02	1.98×10^{-09}
BMI → TC	1.61	0.58	0.96–2.97	0.11	0.02	5.67×10^{-13}
BMI → TG	1.66	1.36	0.91–55.30	0.27	0.02	2.19×10^{-51}
BMI → HDL	−0.20	0.92	−1.71-1.44	−0.20	0.02	4.12×10^{-19}
BMI → LDL	2.01	0.84	1.09–4.31	0.12	0.02	2.21×10^{-11}
	regression adjusting for age (discrete), gender and other health factors			regression adjusting for age (continuous) and gender		
	β	SE	*P*-value	β	SE	*P*-value
BMI → SBP	0.15	0.02	$< 2\times10^{-16}$	0.24	0.01	4.19×10^{-70}
BMI → DBP	0.10	0.02	1.67×10^{-07}	0.22	0.01	1.63×10^{-50}
BMI → FBG	0.07	0.02	1.58×10^{-03}	0.10	0.02	1.66×10^{-09}
BMI → TC	−0.02	0.02	0.23	0.11	0.02	1.87×10^{-12}
BMI → TG	0.09	0.02	1.57×10^{-05}	0.27	0.02	6.24×10^{-51}
BMI → HDL	−0.11	0.03	8.71×10^{-06}	−0.20	0.02	7.45×10^{-19}
BMI → LDL	0.04	0.02	0.02	0.12	0.02	5.12×10^{-11}

no potential causality with HDL (β, − 0.20; 95% CI: − 1.71-1.44). The effect of FBG per 1- kg/m^2 higher BMI was 0.56 (95% CI: − 0.24-2.18). To conclude, the potential causal effects of BMI on SBP, DBP, FBG, TG, TC, HDL and LDL were almost consistent with previous studies. Furthermore, the results showed the robustness of the novel method.

Discussion

In this paper, we present a simple and intuitive method to identify and correct for confounding bias in observation studies using a confounder–exposure nonlinear condition. In cases where the independence assumption between observed confounder and unobserved confounder is violated, a sensible approach shows the almost unbiased causal effect estimation if $P(C_1 = 1) = P(C_2 = 1) = 0.5$. In this sense, our proposed method can be viewed very much as a tool for identifying causal effect in epidemiology.

To identify and estimate causal effect with unobserved confounders, different approaches require different untestable assumptions, such as DID needs an untestable common trend assumption. Fortunately, the approach proposed in this paper requires the nonlinearity assumption of at least two observed confounders $(C_1, C_2, \cdots, C_l)$ and treatment variable X, as well as the conditional expectation of unobserved confounders given observed confounders U is 0. Furthermore, the observed data contain $(C_1, C_2, \cdots, C_l)$ and X which can be utilized to test our nonlinearity assumption. This is exclusive to our method compared with other methods.

For binary variable, the parameter estimates from the LPM can be directly interpreted as the effect of the exposure on the prevalence rate of the outcome which is consistent with the ACE. Conversely, logistic regression or log-linear regression are not applicable in causal inference, since they do not provide a direct interpretation of bivariate associations. However, error terms of LPM with ordinary least square estimation are heteroskedastic and predicted probability can be above 1 or below 0. Therefore, we adopt GMM to deal with these problems. Additionally, our proposed method has no assumption on the independence among observed confounders. This can be widely used in epidemiology to obtain the causal inference of the exposure on the outcome.

Another advantage is the accessibility of two observed confounders satisfying nonlinear condition. Further, we find out that using GMM still fairly reliable in our proposed method if the number of observed confounders is over two.

A number of factors must be considered before implementing the method. First, the independence assumption between observed and unobserved confounders is essential for causal estimate correction. One strategy to overcome this limitation is to select other observed confounders with the probabilities of binary values 0.5. Second, interaction between observed confounders and unobserved confounders on exposure is sufficiently strong. Finally, our method cannot identify the reverse causation. Requiring a priori knowledge or using the idea of Cross-lagged Panel Analysis may avoid this problem.

Identifying causal effects across studies of differing design can therefore prove valuable in further research, whilst agreement with the result of Mendelian Randomization and randomization experiment strengthens confidence in the resulting findings and subsequent inference.

Conclusions

In conclusions, we propose a novel method to control unobserved confounding through double or more binary confounders satisfying a non-linear condition on the exposure which is easy to access. In particular, our method can handle general cases regardless of a continuous or categorical exposure and outcome. Various simulations show better estimation performance by our approach and suggest that our method will be more widely used in observational studies to explore causal association.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12874-020-01049-0.

Additional file 1 : **Appendix A.** Proof of Theorem 1.

Additional file 2 : **Appendix B.** Proof for the equivalence of different choices of $f(\cdot)$ in Eq. (4) for the estimation when the identifiability condition in Theorem 1 holds and Proof for matrix Q^{eff} equals $E[f(C_1, C_2)f(C_1, C_2)^T]$ in Data application results section.

Additional file 3 : **Appendix C.** Simulation parameter settings.

Additional file 4 : **Figure S1.** The Simulation A2 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) C_1 on X, (b) the interaction effect of C_1 and C_2 on X, (c) U on X.

Additional file 5 : **Figure S2.** The Simulation A2 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) X on Y, (b) C_1 on Y, (c) U on Y.

Additional file 6 : **Figure S3.** The Simulation B1 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) C_1 on X, (b) the interaction effect of C_1 and C_2 on X, (c) U on X.

Additional file 7 : **Figure S4.** The Simulation B1 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) X on Y, (b) C_1 on Y, (c) U on Y.

Additional file 8 : **Figure S5** The Simulation B2 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects (a) of C_1 on X, (b) the interaction effect of C_1 and C_2 on X, (c) U on X.

Additional file 9 : **Figure S6.** The Simulation B2 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) X on Y, (b) C_1 on Y, (c) U on Y.

Additional file 10 : **Figure S7.** The Simulation B2 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) the correlation between X and C_1, (b) the correlation between C_1 and C_2.

Additional file 11 : **Figure S8.** The Simulation C1 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects (a) of C_1 on X, (b) the interaction effect of C_1 and C_2 on X, (c) U on X.

Additional file 12 : **Figure S9.** The Simulation C1 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) X on Y, (b) C_1 on Y, (c) U on Y.

Additional file 13 : **Figure S10.** The Simulation C2 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) the correlation between X and C_1, (b) the correlation between C_1 and C_2.

Additional file 14 : **Figure S11.** The Simulation D1 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) C_1 on X, (b) the interaction effect of C_1 and C_2 on X, (c) U on X.

Additional file 15 : **Figure S12.** The Simulation D1 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) X on Y, (b) C_1 on Y, (c) U on Y.

Additional file 16 : **Figure S13.** The Simulation D2 result. Results shows the estimated biases, *SE* and *MSE* from the 3 models for varied effects of (a) the correlation between X and C_1, (b) the correlation between C_1 and C_2.

Abbreviations

IPW: Inverse probability weighing; BMI: Body mass index; SBP: Systolic blood pressure; DBP: Diastolic blood pressure; FBG: Fasting blood glucose; TG: Triglyceride; TC: Total cholesterol; HDL: High density lipoprotein; LDL: Low density lipoprotein; IVs: Instrument variables; IVA: Instrumental variable analysis; DID: Difference-in-differences; RDD: Regression discontinuity design; ACE: Average causal effect; GMM: Generalized moment estimate model; SE: Standard error; MSE: Mean squared error; LPM: Linear probability model

Acknowledgements

We are grateful to the supports of the National Natural Science Foundation of China and Natural Science Foundation of Shandong Province.

Authors' contributions

HL and FX conceived the study. LL, LH, HL contributed to theoretical derivation with assistance from YY, XL and XS. LL, LH and MJ contributed to the data simulation. YY, QW, YX contributed to the application. LL, LH, FY and HL wrote the manuscript with input from all other authors. All authors reviewed and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (No. 81773547) and Natural Science Foundation of Shandong Province (ZR2019ZD02). The funding bodies played no role in the design of the study and collection, analysis, or interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets used and/or analyzed during the present study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

All procedures involving human participants were performed in accordance with the ethical standards of the Ethics Committee of Public Health, Shandong University (No.20190401). Since the study dataset just involved routine available data collection launched by the government, informed consent was not specifically obtained.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Institute for Medical Dataology, Shandong University, 250012 Jinan, Shandong, People's Republic of China. [2]Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, 250012 Jinan, Shandong, People's Republic of China. [3]Synthesis Electronic Technology Co.Ltd, 250012 Jinan, Shandong, People's Republic of China.

Received: 7 February 2020 Accepted: 12 June 2020
Published online: 22 July 2020

References

1. Miao W, Geng Z, Tchetgen Tchetgen EJ. Identifying causal effects with proxy variables of an unmeasured confounder. Biometrika. 2018;105(4):987–93.
2. Trinquart L, Erlinger AL, Petersen JM, Fox M, Galea S. Applying the E value to assess the robustness of epidemiologic fields of inquiry to unmeasured confounding. Am J Epidemiol. 2019;188(6):1174–80.

3. Hjellvik V, De Bruin ML, Samuelsen SO, Karlstad O, Andersen M, Haukka J, Vestergaard P, de Vries F, Furu K. Adjusting for unmeasured confounding using validation data: simplified two-stage calibration for survival and dichotomous outcomes. Stat Med. 2019;38(15):2719–34.
4. Barrowman MA, Peek N, Lambie M, Martin GP, Sperrin M. How unmeasured confounding in a competing risks setting can affect treatment effect estimates in observational studies. BMC Med Res Methodol. 2019;19.
5. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550–60.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.
7. Senn S, Graf E, Caputo A. Stratification for the propensity score compared with linear regression techniques to assess the effect of treatment or exposure. Stat Med. 2007;26(30):5529–44.
8. Heinze G, Juni P. An overview of the objectives of and the approaches to propensity score analyses. Eur Heart J. 2011;32(14):1704–8.
9. He H, Wu P, Chen D-G, SpringerLink (Online service). Statistical causal inferences and their applications in public health research. Cham: Springer International Publishing : Imprint: Springer; 2016.
10. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996;91(434):444–55.
11. Sheehan NA, Didelez V. Commentary: can 'many weak' instruments ever be 'strong'? Int J Epidemiol. 2011;40(3):752–4.
12. Hernan MA, Robins JM. Instruments for causal inference - an epidemiologist's dream? Epidemiology. 2006;17(4):360–72.
13. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Instrumental variables application and limitations. Epidemiology. 2006;17(3):260–7.
14. O'Neill S, Kreif N, Grieve R, Sutton M, Sekhon JS. Estimating causal effects: considering three alternatives to difference-in-differences estimation. Health Serv Outcome Res Methodol. 2016;16:1–21.
15. Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. Annu Rev Public Health. 2018;39:453–69.
16. Thistlethwaite DL, Campbell DT. Regression-discontinuity analysis - an alternative to the ex-post-facto experiment. J Educ Psychol. 1960;51(6):309–17.
17. de la Cuesta B, Imai K. Misunderstandings about the regression discontinuity design in the Study of close elections. Annu Rev Polit Sci. 2016;19:375–96.
18. Lipsitch M, Tchetgen ET, Cohen T. Negative controls a tool for detecting confounding and bias in observational studies. Epidemiology. 2010;21(3):383–8.
19. Lin JH, Lee WC. Complementary log regression for sufficient-cause modeling of epidemiologic data. Sci Rep. 2016;6:39023.
20. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. Stat Methods Med Res. 2017;26(5):2333–55.
21. Hansen LP. Large sample properties of generalized-method of moments estimators. Econometrica. 1982;50(4):1029–54.
22. Moore DF, Tsiatis A. Robust estimation of the variance in moment methods for extra-binomial and extra-poisson variation. Biometrics. 1991;47(2):383–401.
23. Greenland S, Drescher K. Maximum-likelihood-estimation of the attributable fraction from logistic-models. Biometrics. 1993;49(3):865–72.
24. Cameron AC, Trivedi PK. Microeconometrics : methods and applications. Cambridge: Cambridge University Press; 2005.
25. Hansen BE, Horowitz JL. Handbook of econometrics, vol 4 - Engle,RF, McFadden,DL. Economet Theor. 1997;13(1):119–32.
26. Droyvold WB, Midthjell K, Nilsen TI, Holmen J. Change in body mass index and its impact on blood pressure: a prospective population study. Int J Obes. 2005;29(6):650–5.
27. Shamai L, Lurix E, Shen M, Novaro GM, Szomstein S, Rosenthal R, Hernandez AV, Asher CR. Association of body mass index and lipid profiles: evaluation of a broad spectrum of body mass index patients including the morbidly obese. Obes Surg. 2011;21(1):42–7.
28. Siegel D, Meier J, Maas C, Lopez J, Swislocki AL. The effect of body mass index on fasting blood glucose after initiation of thiazide therapy in hypertensive patients. Am J Hypertens. 2008;21(4):438–42.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.