HHS Public Access

Author manuscript

Epidemiology. Author manuscript; available in PMC 2018 March 01.

Published in final edited form as:

Epidemiology. 2017 March ; 28(2): 266–274. doi:10.1097/EDE.0000000000000609.

Parametric mediational g-formula approach to mediation analysis with time-varying exposures, mediators, and confounders

Sheng-Hsuan Lin[1,2,*], **Jessica Young**[1], **Roger Logan**[1], **Eric J. Tchetgen Tchetgen**[1,2], and **Tyler J. VanderWeele**[1,2]

[1]Department of Epidemiology, Harvard School of Public Health, Boston, MA

[2]Department of Biostatistics, Harvard School of Public Health, Boston, MA

Abstract

The assessment of direct and indirect effects with time-varying mediators and confounders is a common but challenging problem, and standard mediation analysis approaches are generally not applicable in this context. The mediational g-formula was recently proposed to address this problem, paired with a semi-parametric estimation approach to evaluate longitudinal mediation effects empirically. In this paper, we develop a parametric estimation approach to the mediational g-formula, including a feasible algorithm implemented in a freely available SAS macro. In the Framingham Heart Study data, we apply this method to estimate the interventional analogues of natural direct and indirect effects of smoking behaviors sustained over a 10-year period on blood pressure when considering weight change as a time-varying mediator. Compared with not smoking, smoking 20 cigarettes per day for 10 years was estimated to increase blood pressure by 1.2 (95 % CI: −0.7, 2.7) mm-Hg. The direct effect was estimated to increase blood pressure by 1.5 (95 % CI: −0.3, 2.9) mm-Hg, and the indirect effect was −0.3 (95% CI: −0.5, −0.1) mm-Hg, which is negative because smoking which is associated with lower weight is associated in turn with lower blood pressure. These results provide evidence that weight change in fact partially conceals the detrimental effects of cigarette smoking on blood pressure. Our work represents, to our knowledge, the first application of the parametric mediational g-formula in an epidemiologic cohort study.

Keywords

mediation analysis; causal inference; time-varying; g-formula; parametrically

[*]**Corresponding**. Sheng-Hsuan Lin, M.D., Department of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA, USA. 677 Huntington Avenue, Boston, MA 02115, USA, Cell: +1 (617) 285-2791, Fax: +1 (617)566-7805, shl517@mail.harvard.edu.

Declarations of competing interest: None.

The mgformula macro is freely accessible with documentation at http://www.hsph.harvard.edu/causal/software/. The dataset is not publicly available.

Introduction

Mediation analysis, a method to decompose the total effect of an exposure on an outcome into direct and indirect effects through a mediator, is essential for investigating pathways or mechanisms in epidemiology and in the social sciences. Causal mediation analysis, defining natural direct and indirect effects based on counterfactual models, extends traditional mediation analysis to settings involving nonlinearities and interactions[1,2]. Numerous methodologic approaches based on causal mediation analysis have been developed in recent years allowing different outcome scales, including additive, multiplicative, and odds ratio scales, as well as other models for time-to-event data[3–8]. Most of the approaches mentioned above only consider a point exposure and a subsequent point mediator. When conducting causal mediation analysis with longitudinal data, the restriction of only one single exposure and mediator neglects exposures or mediators at other time-points, which thus potentially results in loss of valuable information.

Robins has proposed the g-formula to estimate the total effect in settings with time-varying exposures and confounders[9]. In addition, when mediators can be intervened upon, the g-formula can estimate the controlled direct effect by comparing the effects of two different exposure levels and specifying the mediators at certain fixed values. For mediation, however, the natural direct effect and the natural indirect effect involved in effect decomposition are, unfortunately, not identified. VanderWeele and Tchetgen Tchetgen[10] proposed the mediational g-formula to overcome the methodologic challenges of causal mediation analysis with time-varying mediators. As discussed below, this method decomposes the randomized interventional analogue of total effect into interventional analogues of natural direct effect and indirect effect and moves beyond the limitations of a single exposure and a single mediator. Time-varying confounders can also be accounted for provided no unobserved confounding is present. For estimation of the randomized natural direct and indirect effects, VanderWeele and Tchetgen Tchetgen proposed a semi-parametric approach based on inverse probability weighted estimators. As with standard inverse probability weighting, this approach is potentially unstable if the exposure is continuous or the weights are highly variable. Instead, we consider an alternative approach that is potentially more stable and efficient, by implementing a fully parametric mediational g-formula approach, using a user-friendly algorithm implemented in freely available software. We then apply this method to the Framingham Heart Study (FHS) data to investigate the effect of smoking on blood pressure mediated by weight change.

Case study for Framingham dataset: smoking, weight, and blood pressure

In past research, the association between smoking and blood pressure has been controversial[11–16]. In some studies, average blood pressure, as measured using a domestic blood pressure monitor[17–19], is lower among smokers than among non-smokers at particular times of the day[14,20]. According to most literature, among former smokers, smoking cessation is associated with increased blood pressure[21,22] while some studies fail to observe this association[23].

Three possible mechanisms might explain this association. First, smoking activates the autonomic nervous system to increase blood pressure directly, but smoking cessation also

activates the autonomic nervous system, increasing blood pressure among former smokers[24,25]. Second, smoking elevates blood pressure through exacerbating arterial stiffness. Third, smoking decreases blood pressure through weight loss[21]. The adverse effect of smoking on elevated blood pressure might thus be partially concealed by weight loss. Because both smoking status and weight change vary over time, the mediational g-formula can be used to study the extent of this adverse effect mediated by weight loss. In this study, we obtain the estimates by applying our method to the FHS data and demonstrate an application of causal mediation analysis with time-varying mediators.

Methods Development

Notation and review for causal mediation analysis

First consider a setting with an exposure, mediator, and outcome measured at a single time. Let A denote an exposure, Y an outcome, and M a mediator. Let V denote a set of baseline covariates not affected by the exposure. The relationships among these variables are described in Figure 1. Under counterfactual models[26,27], Y_a and M_a denote the counterfactual values of the outcome and the mediator, respectively, if exposure A is set to level a. Y_{am} denotes the counterfactual value of the outcome if exposure A is set to level a, and mediator M is set to level m. In addition, Y_{aMa*} denotes the counterfactual value of the outcome if exposure A is set to level a and mediator M is set to level M_{a*}. Under the consistency assumption[6,28,29], $Y_a = Y$ and $M_a = M$ if $A = a$; $Y_{am} = Y$ if $A = a$ and $M = m$; and the composition assumption that $Y_{aMa*} = Y$ if $A = a$ and $M = M_{a*}$.

Mediation analysis decomposes the overall effect into a direct effect (the effect not through the mediator) and an indirect effect (the effect through the mediator). Under the above counterfactual models, causal mediation analysis usually defines the total effect, natural direct effect, and natural indirect effect to represent the overall, direct, and indirect effects, respectively. Let $A = a$ and $A = a^*$ denote two hypothetical intervention statuses, exposure and non-exposure, respectively. The total effect is defined as $E[Y_a - Y_{a*}]$ or equivalently, $E[Y_{aMa} - Y_{a*Ma*}]$; the natural direct effect is defined as $E[Y_{aMa*} - Y_{a*Ma*}]$, and the natural indirect effect is defined as $E[Y_{aMa} - Y_{aMa*}]$. These effects have been described extensively elsewhere and arguments have been made that they are theoretically appealing for effect decomposition[1,2,6]. However, in the presence of time-varying confounders, the natural direct effect and natural indirect effect are not generally identified from data even if these confounders are observed[30,31]. To address the problem of identification, an alternative definition uses the randomized analogue of total effect, randomized natural direct effect, and randomized natural indirect effect to represent the overall, direct, and indirect effects, respectively[32–34]. Let G_a denote a random draw from the distribution of the mediator amongst those with exposure status $A = a$. When exposure is set to a (or a*), the distribution of mediator among whole population is also determined. Therefore, for every individual, a random draw from this distribution, G_a (or G_{a*}), will be independent of the counterfactual mediator, M_a (or M_{a*}) and outcome (Y_{am}). The randomized analogue of total effect is defined as $E[Y_{aGa}] - E[Y_{a*Ga*}]$. The randomized natural direct effect is defined as $E[Y_{aGa*}] - E[Y_{a*Ga*}]$, which can be interpreted as an effect on the outcome comparing exposure versus no exposure with the mediator in both cases randomly drawn from the distribution of

the population when given no exposure. Finally, the randomized natural direct effect is defined as $E[Y_{aGa}] - E[Y_{aGa^*}]$, interpreted as an effect on the outcome of randomly assigning an individual who is given exposure to a value of the mediator drawn from the distribution of the mediator amongst those given exposure versus not given exposure. We have the decomposition: $E[Y_{aGa}] - E[Y_{a^*Ga^*}] = (E[Y_{aGa}] - E[Y_{aGa^*}]) + (E[Y_{aGa^*}] - E[Y_{a^*Ga^*}])$, so the overall effect decomposes into the sum of the effect through the mediator (i.e. the indirect effect) and the direct effect.

Both natural direct and indirect effects can be identified under four assumptions: (1) $Y_{am} \perp A|V$ (no unmeasured exposure-outcome confounder); (2) $Y_{am} \perp M|V, A$ (no unmeasured mediator-outcome confounder); (3) $M_a \perp A|V$ (no unmeasured exposure-mediator confounder); and (4) $Y_{am} \perp M_{a^*}|V$ (no mediator-outcome confounder affected by exposure)[6,30], when $\perp$ denotes independence and $X \perp Y \mid Z$ denotes that X is independent of Y conditional on Z. Under these assumptions, natural direct effects (NDE) and indirect effects (NIE) are identified by the following expressions:

$$\mathrm{NDE} = \Sigma_v \Sigma_m \{E[Y|a, m, v] - E[Y|a^*, m, v]\} P(m|a^*, v) P(v) \quad (1)$$

$$\mathrm{NIE} = \Sigma_v \Sigma_m E[Y|a, m, v] \{P(m|a, v) - P(m|a^*, v)\} P(v) \quad (2)$$

The fourth assumption holds only under a non-parametric structural equation model and would be violated under several settings. The most common one is the existence of a mediator-outcome confounder L that is affected by exposure (Figure 2), in which case the fourth assumption fails, even if this confounder is observed. When this confounder is a single binary variable, Tchetgen Tchetgen and VanderWeele proposed a method to identify the NDE and NIE by assuming the monotonicity about the effect of exposure on this confounder[35]. A severe shortcoming of this assumption is that even if the mediator is restricted to occurring immediately after the exposure, the assumption cannot be ensured. As a result, this approach is not generally applicable for settings with time-varying mediators. However, even if this assumption fails, given assumptions (1) to (3) hold and the mediator-outcome confounder is observed, the randomized natural direct effect (rNDE) and randomized natural indirect effect (rNIE) in the second definition are still identifiable from the data by the empirical expressions given by[34]:

$$\mathrm{rNDE} = \Sigma_{v,l,m} \{E[Y|a, l, m, v] P(1|a, v) - E[Y|a^*, l, m, v] P(1|a^*, v)\} P(m|a^*, v) P(v) \quad (3)$$

$$\mathrm{rNIE} = \Sigma_{v,l,m} E[Y|a, l, m, v] P(l|a, v) \{P[m|a, v] - P(m|a^*, v)\} P(v) \quad (4)$$

To understand the difference in the effects better, note that the natural direct effect and natural indirect effect cannot be checked by designing a randomized trial even if we were

able to intervene on both the exposure and the mediator. However, it is possible to design a two-stage trial to estimate randomized natural direct and indirect effects. A randomized trial could first be conducted to obtain the empirical distribution of counterfactual mediator given exposed and non-exposed by randomizing the exposure and measuring the mediator distributions. We can then estimate randomized natural indirect effect in a second trial by the effect on the outcome of assigning an individual who is given the exposure to a value of the mediator sampled from the marginal distribution of the mediator amongst those given exposure versus no exposure, using the empirical distributions of the mediator estimated from the first randomized trial. Similarly, we can estimate randomized natural direct effect by a direct effect comparing exposure versus no exposure with the mediator in both cases randomly drawn from the empirical distribution of the population when given no exposure, which was also estimated from the first randomized trial.

Notation and review of mediation analysis with time-varying mediators and the mediational g-formula

Consider exposures, mediators, and confounders that vary over time in longitudinal settings with T measurements at time t = 0, 1, 2, …, T−1. Let (A(0), A(1), …, A(T−1)), (M(0), M(1), …, M(T−1)), and (L(0), …, L(T−1)) denote values of the time-varying exposures, mediators, and confounders at periods 1, …, T, with the final outcome of interest Y. The initial baseline confounders are included in L(0). Figure 3 depicts a possible data generating mechanism under which these assumptions would hold.

For any variable W, let $\overline{\mathrm{W(t)}}=(\mathrm{W}(0),\mathrm{W}(1),\ldots,\mathrm{W}(\mathrm{t}))$ and let $\overline{\mathrm{W}}=\overline{\mathrm{W(T-1)}}=(\mathrm{W}(0),\mathrm{W}(1),\ldots,\mathrm{W}(\mathrm{T}-1))$. Let $\mathrm{Y}_{\overline{\mathrm{am}}}$ be the counterfactual value of Y given $\bar{\mathrm{A}}$ is set to $\bar{\mathrm{a}}$ and $\bar{\mathrm{M}}$ is set to $\bar{\mathrm{m}}$. Let $\mathrm{M}_{\bar{\mathrm{a}}}(\mathrm{t})$ be the counterfactual value of M(t) given $\overline{\mathrm{A(t)}}$ is set to $\overline{\mathrm{a(t)}}$. Let $\mathrm{G}_{\bar{\mathrm{a}}}(\mathrm{t})$ denote a random draw from the distribution of the mediator $\mathrm{M}_{\bar{\mathrm{a}}}(\mathrm{t})$. Let $\bar{\mathrm{A}}=\bar{\mathrm{a}}$ and $\bar{\mathrm{A}}=\overline{\mathrm{a}^*}$ denote two hypothetical intervention statuses, for example, exposed from t = 0 to T−1 and non-exposed from t = 0 to T−1, respectively. In this setting, we define total effect as $\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}}]-\mathrm{E}[\mathrm{Y}_{\overline{\mathrm{a}*}}]$ (i.e. $\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\overline{\mathrm{M}}\bar{\mathrm{a}}}]-\mathrm{E}[\mathrm{Y}_{\overline{\mathrm{a}*}\overline{\mathrm{M}}\overline{\mathrm{a}*}}]$), natural direct effect, as $\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\overline{\mathrm{M}}\overline{\mathrm{a}*}}]-\mathrm{E}[\mathrm{Y}_{\overline{\mathrm{a}*}\overline{\mathrm{M}}\overline{\mathrm{a}*}}]$, and natural indirect effect $\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\overline{\mathrm{M}}\bar{\mathrm{a}}}]-\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\overline{\mathrm{M}}\overline{\mathrm{a}*}}]$; we aslo define randomized analogues of total effect as $\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\bar{\mathrm{G}}\bar{\mathrm{a}}}]-\mathrm{E}[\mathrm{Y}_{\overline{\mathrm{a}*}\bar{\mathrm{G}}\overline{\mathrm{a}*}}]$, randomized analogue of natural direct effect as $\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\bar{\mathrm{G}}\overline{\mathrm{a}*}}]-\mathrm{E}[\mathrm{Y}_{\overline{\mathrm{a}*}\bar{\mathrm{G}}\overline{\mathrm{a}*}}]$, and randomized analogue of natural indirect effect as $\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\bar{\mathrm{G}}\bar{\mathrm{a}}}]-\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\bar{\mathrm{G}}\overline{\mathrm{a}*}}]$. We can decompose the total effect into the natural direct effect and natural indirect effect. Similarly, randomized analogue of total effect is decomposed into randomized analogues of natural direct effect and of natural indirect effect (i.e., $\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\bar{\mathrm{G}}\bar{\mathrm{a}}}]-\mathrm{E}[\mathrm{Y}_{\overline{\mathrm{a}*}\bar{\mathrm{G}}\overline{\mathrm{a}*}}]=\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\bar{\mathrm{G}}\overline{\mathrm{a}*}}]-\mathrm{E}[\mathrm{Y}_{\overline{\mathrm{a}*}\bar{\mathrm{G}}\overline{\mathrm{a}*}}]+\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\bar{\mathrm{G}}\bar{\mathrm{a}}}]-\mathrm{E}[\mathrm{Y}_{\bar{\mathrm{a}}\bar{\mathrm{G}}\overline{\mathrm{a}*}}]$.

If the entire vector $\bar{\mathrm{A}}=\overline{\mathrm{A(T-1)}}$ is taken as the joint exposure of interest and $\overline{\mathrm{M}}=\overline{\mathrm{M(T-1)}}$ as the mediators of interest, then assumption (4) is violated because the variable L(1) is affected by A(0) and confounds the mediator-outcome relationship between M(1) and Y (similarly, the L(t) is affected by $\overline{\mathrm{A(t-1)}}$ and confounds the relationship between M(t) and Y, when t = 1, …, T−1). Therefore, natural direct and indirect effects cannot be identified in this setting. However, randomized natural direct and indirect effects are still identifiable

under the following three assumptions for all t: (1) $Y_{\overline{am}^{\perp}} A(t) | \overline{A(t-1)}, \overline{M(t-1)}, \overline{L(t)}$ (no exposure–outcome confounding conditional on the past variables), (2) $Y_{\overline{am}^{\perp}} M(t) | \overline{A(t)}, \overline{M(t-1)}, \overline{L(t)}$ (no mediator–outcome confounding conditional on the past variables), and (3) $M_{\bar{a}}(t) \perp A(t) | \overline{A(t-1)}, \overline{M(t-1)}, \overline{L(t)}$ (no exposure–mediator confounding conditional on the past variables)[10]. Given the three assumptions, VanderWeele and Tchetgen Tchetgen[10] show that the randomized natural direct (rNDE) and indirect effects (rNIE) are identified non-parametrically by the following equations:

$$\mathrm{rNDE} = Q(\bar{a}, \overline{a^*}) - Q(\overline{a^*}, \overline{a^*}) \quad (5)$$

$$\mathrm{rNIE} = Q(\bar{a}, \bar{a}) - Q(\bar{a}, \overline{a^*}) \quad (6)$$

$$\text{where } Q(\bar{a}_1, \bar{a}_2) = \sum_{\overline{m}} \sum_{\bar{l}} E[Y | \overline{a_1}, \overline{m}, \bar{l}] \prod_{t=0}^{T-1} P(l(t) | \overline{a_1(t-1)}, \overline{m(t-1)}, \overline{l(t-1)}) \times \sum_{\overline{l'}} \prod_{t=0}^{T-1} P(m(t) | \overline{a_2(t)}, \overline{m(t-1)}, \overline{l'(t)}) P(l'(t) | \overline{a_2(t-1)}, \overline{m(t-1)}, \overline{l(t-1)}) \quad (7)$$

As in VanderWeele and Tchetgen Tchetgen, we refer to this expression $Q(\bar{a}_1, \bar{a}_2)$ above as the mediational g-formula. When there are no mediators, i.e. $\bar{M}$ equals empty, this formula (7) reduces to the standard g-formula:

$$Q(\bar{a}) = \sum_{\bar{l}} E[Y | \overline{a_1}, \bar{l}] \prod_{t=0}^{T-1} P(l(t) | \overline{a_1(t-1)}, \overline{l(t-1)}) \quad (8)$$

Further, rNDE (5) and rNIE (6) reduce to the natural direct effect and natural indirect effect, respectively, when there are no time-varying confounders[10].

Parametric mediational g-formula

VanderWeele and Tchetgen Tchetgen described how to use inverse probability weighting of marginal structural models to estimate the mediational g-formula $Q(\bar{a}_1, \bar{a}_2)$ (7) in realistic high-dimensional settings[10]. However, this approach can perform poorly with continuous exposures and mediators and can also be inefficient. As an alternative, an adaptation of the standard parametric g-formula[9,36] can be used to parametrically estimate the mediational g-formula in high-dimensional settings and, in turn, the randomized natural direct effect (5) and randomized natural indirect effect (6).

We begin by briefly reviewing the standard parametric g-formula. This approach parametrically estimates the standard g-formula by (i) fitting parametric models for the joint density of the outcome and time-varying covariates and (ii) using the estimated parameters of these models to simulate many covariate histories consistent with the exposure intervention $\bar{a}$. Specifically, the following algorithm, which is implemented in a publicly available SAS macro[37], can be used to parametrically estimate the standard g-formula $Q(\bar{a})$:

1. Fit parametric models for the observed data:
   - **(1a)** For times $t \geq 0$, fit parametric models for the joint density of the confounders and exposures at t given the measured past.
   - **(1b)** Fit a parametric model for the mean of the outcome at the end of follow-up given the measured past.
2. Set baseline confounders and exposures to the observed sample values. Recursively, for each subject i = 1, …, n and for each time t = 0, 1, 2, …, T−1:
   - **(2a)** For $t \geq 0$, generate time t confounders and exposures based on the estimated model coefficients of (1a) and previously generated exposures and confounders under intervention.
   - **(2b)** Assign time t exposures under intervention $\bar{a}$.
3. Simulate the outcome for each of the n generated histories in step 2 based on the estimated model coefficients of (1b).
4. Take the mean over n simulated outcomes in (3) to estimate $Q(\bar{a})$.

Here, we have adapted the above algorithm and associated SAS macro code to parametrically estimate the mediational g-formula $Q(\bar{a}_1, \bar{a}_2)$, the randomized analogues of total effect, of natural direct effect, and of natural indirect effect. The primary difference between the parametric g-formula and the parametric mediational g-formula is, under the latter algorithm, the estimated model coefficients from step (1) are additionally used to estimate the joint distribution of the time-varying mediators (marginal over all other covariates) under both exposure interventions $\bar{a}$ and $\overline{a^*}$. These are then used to assign values of the mediator under the joint exposure and mediator interventions $(\bar{a},\bar{G}_{\bar{a}})$, $(\bar{a}, \bar{G}_{\overline{a^*}})$, $(\overline{a^*}, \bar{G}_{\bar{a}})$ and $(\overline{a^*}, \bar{G}_{\overline{a^*}})$. The algorithm is as follows:

1. Fit parametric models for the observed data:
   - **(1a)** For times $t \geq 0$, fit parametric models for the joint density of the confounders, exposures and mediators at t given the measured past.
   - **(1b)** Fit a parametric model for the mean of the outcome at the end of follow-up given the measured past.
2. Estimate the joint distribution of time-varying mediators under time-varying exposure interventions $\bar{a}$ and $\overline{a^*}$:
   - **(2a)** Set baseline (t = 0) covariates to the observed values for subject i. Recursively, for each time t = 0, …, T−1 and each subject i = 1, …, n:

(2a.i) For t ≥ 0, generate time t confounders, exposure, and mediator based on the estimated model coefficients of (1a) and previously generated covariates under the time-varying exposure intervention ā through t−1.

(2a.ii) Assign time t exposure under the intervention ā.

(2b) For each t = 0,…T−1, randomly permute the n values of the joint mediators assigned under intervention ā in (2a). For each t, save this permutation for use in (3) below (we obtain $\bar{G}_{\bar{a}}$ in this step).

(2c) Repeat (2a) replacing intervention ā with $\overline{a^*}$.

(2d) Repeat (2b) replacing intervention ā with $\overline{a^*}$ (we obtain $\bar{G}_{\overline{a^*}}$ in this step).

3. Estimate Q(ā, ā), $Q(\bar{a}, \overline{a^*})$, $Q(\overline{a^*}, \bar{a})$ and $Q(\overline{a^*}, \overline{a^*})$ by repeating the following for each $(\overline{a_1}, \overline{a_2})=(\bar{a}, \bar{a})$, $(\bar{a}, \overline{a^*})$, $(\overline{a^*}, \bar{a})$ and $(\overline{a^*}, \overline{a^*})$:

(3a) Recursively for each time t = 0,…,T−1 and each subject i = 1,…,n:

(3a.i) Repeat (2a.i) but replacing “time-varying exposure intervention ā through t−1” with the joint “time-varying exposure and mediator intervention $(\overline{a_1}, \bar{G}_{\overline{a_2}})$”.

(3a.ii) Assign the time t mediator as the i^{th} component of the permuted vector for time t from (2b) $(\text{if } \overline{a_2}=\bar{a})$ or (2d) $(\text{if } \overline{a_2}=\overline{a^*})$.

(3a.iii) Assign time t exposure under the intervention $\overline{a_1}$.

(3b) Simulate the outcome given each of the i = 1,…,n histories based on the estimated model coefficients of (1b) and the histories generated in (3a).

(3c) Estimate $Q'(\overline{a_1}, \overline{a_2})$ as the mean over the n simulated outcomes in (3b).

(3d) Repeat (1) to (3c) for some fixed number K (e.g. 25) times, using different permutation in (2b) for each time.

(3e) Estimate $Q(\overline{a_1}, \overline{a_2})$ as the mean of the K (e.g. 25) values of $Q'(\overline{a_1}, \overline{a_2})$ in (3d).

The algorithm can stop at (3c) and use the $Q'(\overline{a_1}, \overline{a_2})$ as the unbiased estimate of $Q(\overline{a_1}, \overline{a_2})$. However, the repeated steps in (3d) can improve standard errors for smaller sample sizes. Estimates of the randomized natural direct and indirect effects are then calculated from the estimates of the four $Q(\overline{a_1}, \overline{a_2})$ in (3). 95% confidence intervals are calculated based on repeating the above algorithm in 500 bootstrap samples of the original n observations. This algorithm can be implemented with the mgformula macro, freely accessible with documentation at http://www.hsph.harvard.edu/causal/software/. However, the data are not available on the website. Please see the eAppendix for details.

Data Application—Beginning in 1948 in Framingham, Massachusetts, the FHS is a longitudinal cohort study. The original cohort consists of 5,209 participants aged from 30 to 62 years old without cardiovascular disease (CVD) history at baseline. All the participants underwent examinations at the beginning of the study and routinely every two years after that. During each exam, potential CVD risk factors were collected, including socio-demographic data, lifestyle characteristics, detailed medical history, physical examination data, and blood samples. Further details on the design of FHS have been described elsewhere[23,38]. All participants had provided written informed consent and the protocol was approved by the Institutional Review Board at Boston University Medical Center. The purpose of the analysis here is to illustrate the parametric meditational g-formula approach and software.

We specify exam 3 as the first exam and exams 1 and 2 as pre-baseline covariates to allow the function of the past in the models of step (1) of the estimation algorithm to depend on two lagged periods of the covariates. We follow the cohort for ten years (i.e. five visits) to reduce the proportion of death or loss to follow up to limit selection bias by death. Four exclusion criteria are listed below: (1) death or loss to follow up during the period before exam 7 (the end of follow-up); (2) no record at baseline on weight, height, smoking status, former smoking history, systolic blood pressure (SBP), or total cholesterol; (3) diagnosis of diabetes, cancer, or CVD at baseline; and (4) missing value for smoking status or BMI missing more than once. After these exclusions, 3,116 participants remain eligible for our analysis. For simplicity, we now refer to the original FHS exams 3,…, 7 as exams 1,…, 5.

SBP at exam 5 is the outcome Y. BMI during follow-up (exam 1 to 5) is the mediator $\bar{M}$. The exposure $\bar{A}$ is smoking status during follow up, measured as self-reported average number of cigarettes smoked per day. For missing BMI value or smoking status at a single time period, we carry forward the last observed value/status for one exam period only. We consider "smoking 20 cigarettes per day" and "no smoking" during follow-up as two hypothetical intervention levels $\bar{A} = \bar{a}$ and $\bar{A}=\overline{a^*}$. Time-varying covariates $\bar{L}$ include the exam number, the systolic blood pressure (mm-Hg) at last exam, total non-fasting cholesterol level (mg/dl), and the usage of antihypertensive drugs. Baseline covariates include gender, age (years), height (meters), education (≤ 8th grade, some high school, high school graduate, some college, college graduate, post-graduate), occupation before retirement (executive/supervisory, technical, laborer, clerical, sales, housewife), marital status (single, married, widowed/divorced), and tobacco use at baseline (never user, current user, and past user). All covariates and the corresponding models are listed in Table 1.

The parametric g-formula is used to estimate the total effect of smoking 20 cigarettes per day v.s. no smoking on SBP and on BMI at exam 5, by the g-formula macro. The parametric mediational g-formula is applied to conduct mediation analysis with time-varying mediators and exposures by our newly developed mgformula macro. We specify models for the outcome mean, as well as for each time-varying covariate (including mediator, exposure, and confounders) at each time point. We use current covariates and covariates at one period back (one lagged model) as the predictors. Specifically, we regress Y on main effects for A(5), M(5), L(5), A(4), M(4), and L(4). For t = 0, 1, 2, …, 5, we regress L(t) on A(t−1), M(t−1), and L(t−1); regress M(t) on A(t), A(t−1), M(t−1), L(t), and L(t−1); and regress A(t) on A(t

−1), M(t−1), L(t), and L(t−1) (please refer to the eAppendix for more details). All analyses are conducted using SAS 9.4 (Cary, NC).

Results

The demographic and baseline health characteristics for smokers (n = 1,759), non-smokers (n = 1,174), and quitters (n = 183) are shown in Table 2. Compared with non-smokers, smokers have higher male proportion, younger age, and better education level. The majority of non-smokers are female (84%) and half of them have the occupation of housewife. The occupations of the smokers are mainly supervisor, laborer, and housewife. At baseline, the smokers appear to have better health status for lower systolic blood pressure, cholesterol level, and BMI.

We use the g-formula to estimate the total effect of smoking (20 cigarettes per day v.s. no smoking for 10 years) on SBP and BMI in Table 3. Smoking elevates SBP by 1.2 mm-Hg and reduces BMI by 0.2 kg/m^2. In Table 4, we use the parametric mediational g-formula to simulate systolic blood pressure at the end of 10-year follow-up under no smoking with BMI distributed as the BMI under no smoking, smoking 20 cigarettes per day with BMI distributed as the BMI under smoking 20 cigarettes per day, no smoking with BMI distributed as the BMI under smoking 20 cigarettes per day, and smoking 20 cigarettes per day with BMI distributed as the BMI under no smoking. We then estimate the randomized total effect, randomized natural direct effects, and randomized natural indirect effects of smoking on systolic blood pressure mediated by BMI. The estimate of randomized analogue of total effect is 1.2 (95 % CI: −0.7, 2.7) mm-Hg, the estimate of randomized analogue of natural direct effect is 1.5 (95 % CI: −0.3, 2.9) mm-Hg, and the estimate of randomized analogue of natural indirect effect is −0.3 (95 % CI: −0.5, −0.1) mm-Hg. The directions of randomized natural direct and indirect effects are different, suggesting that the seemingly protective mediated effect of smoking through BMI may partially mask the detrimental direct effect of smoking on systolic blood pressure.

Discussion

To our knowledge, this is the first paper to provide a fully parametric method for causal mediation analysis with time-varying mediators. We develop an algorithm and corresponding SAS macro to use the mediational g-formula parametrically. We use the parametric approach to obtain estimates for the mediational g-formula by adapting the g-formula macro externally to build our SAS macro. Similar to the g-formula, we use Monte-Carlo simulation and bootstrapping methods for point and interval estimation, respectively. Since the estimation is an approximation of the maximum likelihood estimation, this estimation is asymptotically efficient provided the regression models are all correctly specified, while the inverse probability weighting does not achieve the efficiency bound in a model where parametric assumptions about the weights and the marginal structural modeling are correct. In addition, like other simulation-based methods[39], the approach here has the advantage of allowing for very flexible models such as quadratic linear models.

The parametric mediational g-formula provides a powerful method to investigate the mechanisms of an effect through time-varying mediators. Traditional techniques, allowing only one observation of the mediator and restricting it to variables occurring immediately after the exposure, inadequately capture the indirect effect when the exposures affect the mediators over time. One alternative approach is to estimate the controlled direct effect, the effect of the exposures on the outcome when fixing the mediators at certain values. Obtained by applying the g-formula and specifying the mediators to these values, the controlled direct effect estimate provides valuable information for policy makers. For example, from FHS data we can provide the effect of smoking cessation decreasing systolic blood pressure if BMI is fixed to a certain level. This is also described elaborately elsewhere[40]. The effects from the meditational g-formula will, however, further allow for effect decomposition.

This study provides direct evidence for the hypothesis that "the adverse effect of smoking on high blood pressure is partially concealed by weight loss"[13,21]. The concealment of the effect is partial because smoking is not substantially associated with weight loss (Table 3). Some studies report that smoking cessation increases blood pressure[21,22]. The discrepancy with these studies might be attributed to a different analysis approach or simply the special characteristics of FHS participants.

Several limitations of this study should be noted. First, the use of analogue of total effect, randomized natural direct effect, and randomized natural indirect effect, defined based on the stochastic interventions, results in different interpretations from the total effect, natural direct effect, and natural indirect effect[34]. This is inevitable for causal mediation analysis with time-varying mediators because the natural direct and indirect effect are not generally identified by empirical data in the presence of time-varying confounders, randomized natural direct and indirect effect can still serve as an analogue of natural direct and indirect effect, and have the advantage that they can be verified by randomized controlled trial, while natural direct and indirect effect cannot. Second, in the macro we have developed, the outcome can only be affected by the covariates at the most recent three exams because of a similar restriction of the g-formula macro employed for continuous/binary outcome types. This is not a limitation of the methodology itself, but only of the current implementation. The earlier covariates affect the outcome only through these recent covariates. We specify the outcome variables in previous exams (i.e. the previous systolic blood pressure values) as the time-varying covariates. Thus, the earlier covariates can also affect the outcome through the previous outcome variables. Third, we have not adjusted for selection bias here. Selection bias or truncation by death is a difficult problem for causal inference generally and for mediation analysis, even with only one single mediator. Most of the literature makes a sequential ignorability assumption that survival is effectively randomized conditional on the past. Under this assumption, the result can be interpreted as what would happen to the population if one could intervene to prevent death for everyone. Alternatively, one could also pursue sensitivity analysis approaches and there are also alternative stronger assumptions that would allow one to interpret the results without necessarily intervening upon death[41,42]. We hope to address this in future work but the extension of this to time-varying exposures and mediators is substantial. For the purposes of the illustration, we have simply focused on complete-case data and restricted our follow-up to ten years to partially address this selection bias. The relatively low proportion of loss to follow-up (< 20%) perhaps partially

mitigates this problem. A general disadvantage of our proposed approach is that it may be particularly prone to bias due to model misspecification. Some misspecification may be theoretically guaranteed in complex longitudinal settings as considered here when the null hypothesis of no causal time-varying treatment effect is true, a problem known as the g-null paradox[43]. The magnitude of bias implied by the g-null paradox is not guaranteed and, depending on the setting, may be large or small. For some further consideration of the g-null paradox using numerical examples, see Young and Tchetgen Tchtegen[44]. The aforementioned inverse probability weighting estimator offers an alternative method which is not subject to the g-null paradox and might be used for the population estimands considered here[45]. Finally, the analysis is subject to potential violation of the confounding assumptions. Future research could develop sensitivity analysis techniques for violations of these assumptions for our method.

Conclusion

The parametric mediational g-formula serves as a powerful and useful tool for mediation analysis with longitudinal data. Researchers can apply our method to disentangle the complicated causal mechanisms arising from time-varying mediators, exposures, and confounders. Further issues concerning the interpretation of the interventional direct and indirect effects can be found in VanderWeele and Tchetgen Tchetgen[10]. Using this method, we provide evidence that weight change in fact partially conceals the detrimental effect of cigarette smoking on blood pressure.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Sources of financial support: None

References

1. Pearl, J. Proceedings of the seventeenth conference on uncertainty in artificial intelligence. Morgan Kaufmann Publishers, Inc; 2001. Direct and indirect effects; p. 411-420.
2. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology. 1992:143–155. [PubMed: 1576220]
3. Tchetgen EJT, Shpitser I. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness and sensitivity analysis. The Annals of Statistics. 2012; 40(3):1816–1845. [PubMed: 26770002]
4. Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. Psychological methods. 2013; 18(2):137. [PubMed: 23379553]
5. van der Laan MJ, Petersen ML. Direct effect models. The international journal of biostatistics. 2008; 4(1)
6. VanderWeele T, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. Statistics and its Interface. 2009; 2:457–468.
7. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology (Cambridge, Mass.). 2010; 21(4):540.

8. VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. American journal of epidemiology. 2010; 172(12):1339–1348. [PubMed: 21036955]
9. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling. 1986; 7(9): 1393–1512.
10. VanderWeele TJ, Tchetgen Tchetgen E. Mediation Analysis with Time-Varying Exposures and Mediators. 2014
11. Chu N-F, Ding Y-A, Wang D-J, Shieh S-M. Relationship between smoking status and cardiovascular disease risk factors in young adult males in Taiwan. Journal of cardiovascular Risk. 1996; 3(2):205–208. [PubMed: 8836864]
12. Gordon T, Kannel WB, Dawber TR, McGee D. Changes associated with quitting cigarette smoking: the Framingham Study. American heart journal. 1975; 90(3):322–328. [PubMed: 1163424]
13. John U, Meyer C, Hanke M, Völzke H, Schumann A. Smoking status, obesity and hypertension in a general population sample: a cross-sectional study. Qjm. 2006; 99(6):407–415. [PubMed: 16687420]
14. Okubo Y, Suwazono Y, Kobayashi E, Nogawa K. An association between smoking habits and blood pressure in normotensive Japanese men: a 5-year follow-up study. Drug and alcohol dependence. 2004; 73(2):167–174. [PubMed: 14725956]
15. Radi S, Lang T, Lauwers-Cances V, Chatellier G, Fauvel J, Larabi L, De Gaudemaris R. One-year hypertension incidence and its predictors in a working population: the IHPAF study. Journal of human hypertension. 2004; 18(7):487–494. [PubMed: 14961044]
16. Retnakaran R, Hanley AJ, Connelly PW, Harris SB, Zinman B. Cigarette smoking and cardiovascular risk factors among Aboriginal Canadian youths. Canadian Medical Association Journal. 2005; 173(8):885–889. [PubMed: 16217111]
17. Mann SJ, James GD, Wang RS, Pickering TG. Elevation of ambulatory systolic blood pressure in hypertensive smokers: a case-control study. Jama. 1991; 265(17):2226–2228. [PubMed: 2013955]
18. Verdecchia P, SchillacI G, Borgioni C, Ciucci A, Zampi I, Battistelli M, Gattobigio R, Sacchi N, Porcellati C. Cigarette smoking, ambulatory blood pressure and cardiac hypertrophy in essential hypertension. Journal of hypertension. 1995; 13(10):1209–1216. [PubMed: 8586813]
19. Gambini G, Di Cato L, Pinchi G, Valori C. 24-hour ambulatory monitoring of arterial blood pressure and the sympathetic nervous system in hypertensive smokers. Giornale italiano di cardiologia. 1997; 27(11):1153–1157. [PubMed: 9463059]
20. Berglund G, Wilhelmsen L. Factors related to blood pressure in a general population sample of Swedish men. Acta medica scandinavica. 1975; 198(1–6):291–298. [PubMed: 1189986]
21. Janzon E, Hedblad B, Berglund G, Engström G. Changes in blood pressure and body weight following smoking cessation in women. Journal of internal medicine. 2004; 255(2):266–272. [PubMed: 14746564]
22. Lee D-H, Ha M-H, Kim J-R, Jacobs DR. Effects of smoking cessation on changes in blood pressure and incidence of hypertension a 4-year follow-up study. Hypertension. 2001; 37(2):194–198. [PubMed: 11230270]
23. Dawber TR, Meadors GF, Moore FE Jr. Epidemiological Approaches to Heart Disease: The Framingham Study*. American Journal of Public Health and the Nations Health. 1951; 41(3):279–286.
24. Cryer PE, Haymond MW, Santiago JV, Shah SD. Norepinephrine and epinephrine release and adrenergic mediation of smoking-associated hemodynamic and metabolic events. New England journal of medicine. 1976; 295(11):573–577. [PubMed: 950972]
25. Grassi G, Seravalle G, Calhoun DA, Bolla GB, Giannattasio C, Marabini M, Del Bo A, Mancia G. Mechanisms responsible for sympathetic activation by cigarette smoking in humans. Circulation. 1994; 90(1):248–253. [PubMed: 8026005]
26. Rubin DB. Formal mode of statistical inference for causal effects. Journal of Statistical Planning and Inference. 1990; 25(3):279–292.
27. Hernán M. A definition of causal effect for epidemiological research. Journal of Epidemiology and Community Health. 2004; 58(4):265–271. [PubMed: 15026432]

28. Pearl J. Causal inference in statistics: An overview. Statistics Surveys. 2009; 3:96–146.
29. VanderWeele TJ. Concerning the consistency assumption in causal inference. Epidemiology. 2009; 20(6):880–883. [PubMed: 19829187]
30. Avin, C., Shpitser, I., Pearl, J. Identifiability of path-specific effects. Department of Statistics, UCLA; 2005.
31. VanderWeele, T. Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press; 2015.
32. Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2007; 69(2):199–215.
33. Didelez V, Dawid P, Geneletti S. Direct and indirect effects of sequential treatments. arXiv preprint arXiv:1206.6840. 2012
34. VanderWeele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiology. 2014; 25(2):300–306. [PubMed: 24487213]
35. Tchetgen EJT, VanderWeele TJ. On identification of natural direct effects when a confounder of the mediator is directly affected by exposure. Epidemiology (Cambridge, Mass.). 2014; 25(2):282.
36. Taubman SL, Robins JM, Mittleman MA, Hernán MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. International Journal of Epidemiology. 2009; 38(6):1599–1611. [PubMed: 19389875]
37. Roger Logan ST, Young Jessica, Picciotto Sally, Hernán Miguel A. GFORMULA SAS MACRO - Estimates the mean of a dichotomous outcome at end of follow-up under general interventions on time-varying treatments in observational studies using the parametric g-formula. 2015 http://www.hsph.harvard.edu/causal/software/.
38. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J. Factors of Risk in the Development of Coronary Heart Disease—Six-Year Follow-up Experience The Framingham Study. Annals of internal medicine. 1961; 55(1):33–50. [PubMed: 13751193]
39. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychological methods. 2010; 15(4):309. [PubMed: 20954780]
40. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. Epidemiology. 2009; 20(1):18–26. [PubMed: 19234398]
41. Tchetgen Tchetgen EJ. Identification and estimation of survivor average causal effects. Statistics in medicine. 2014; 33(21):3601–3628. [PubMed: 24889022]
42. Tchetgen EJT, Glymour MM, Shpitser I, Weuve J. Rejoinder: to weight or not to weight?: on the relation between inverse-probability weighting and principal stratification for truncation by death. Epidemiology. 2012; 23(1):132–137.
43. Robins, JM., Wasserman, L. Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. Proceedings of the thirteenth conference on uncertainty in artificial intelligence; Morgan Kaufmann Publishers Inc; 1997. p. 409-420.
44. Young JG, Tchetgen Tchetgen EJ. Simulation from a known Cox MSM using standard parametric models for the g-formula. Statistics in medicine. 2014; 33(6):1001–1014. [PubMed: 24151138]
45. Van der Laan, MJ., Rose, S. Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media; 2011.

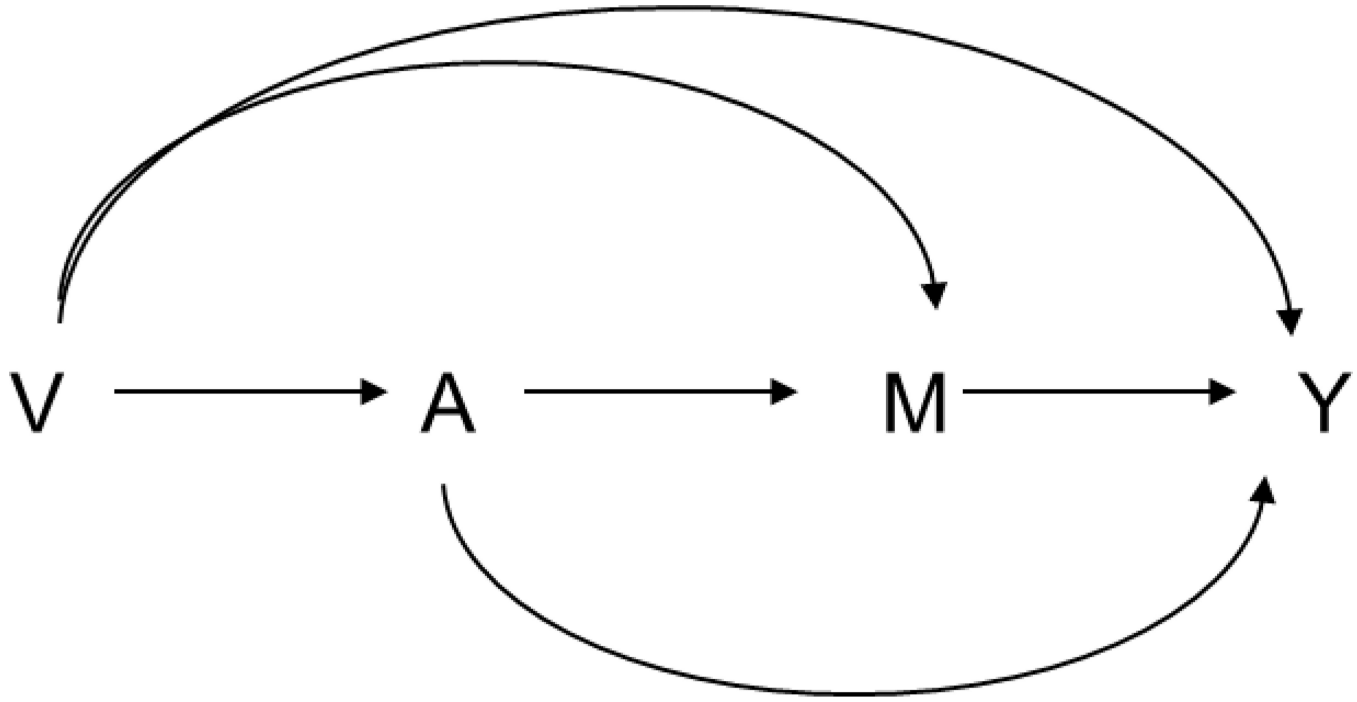

Figure 1.
Simple model for mediation analysis.

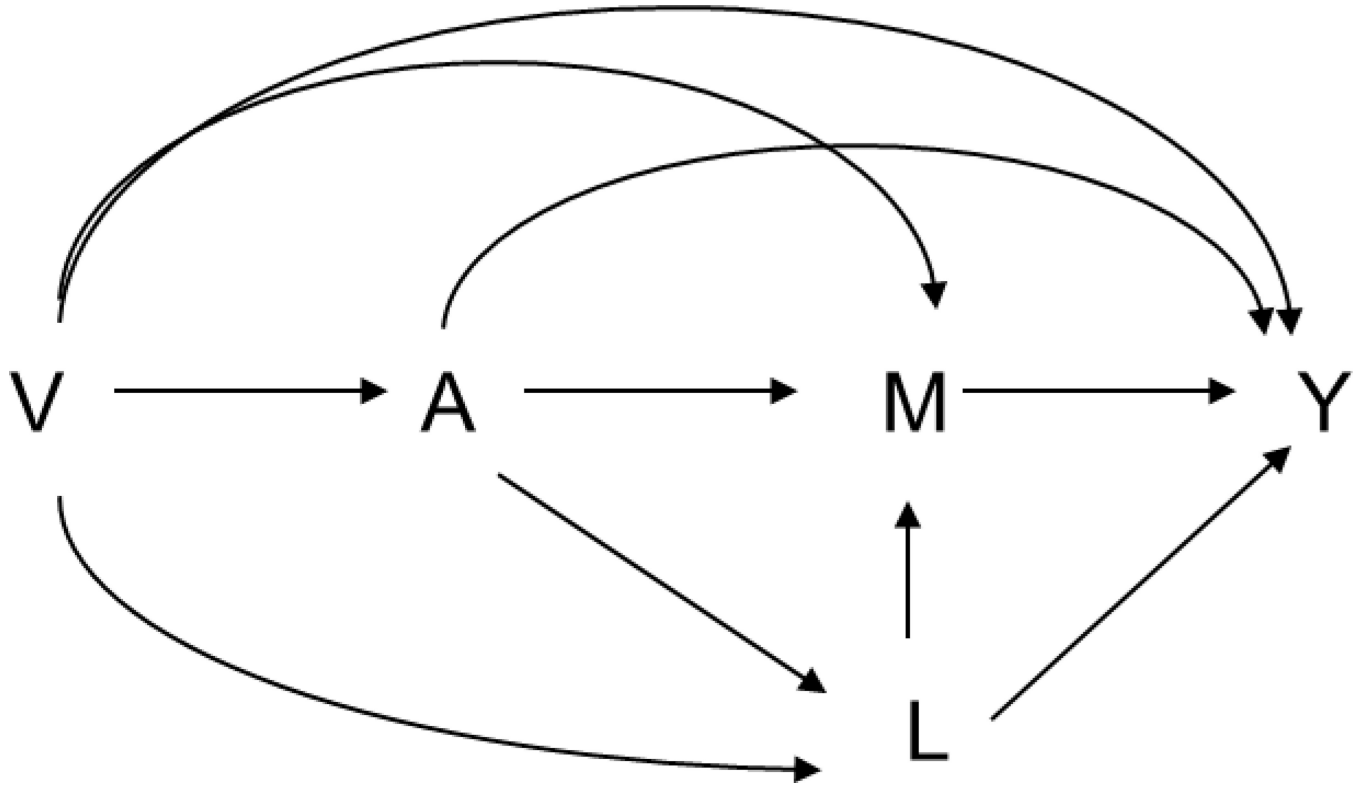

Figure 2.
Mediation analysis with a mediator-outcome confounder L that is affected by exposure.

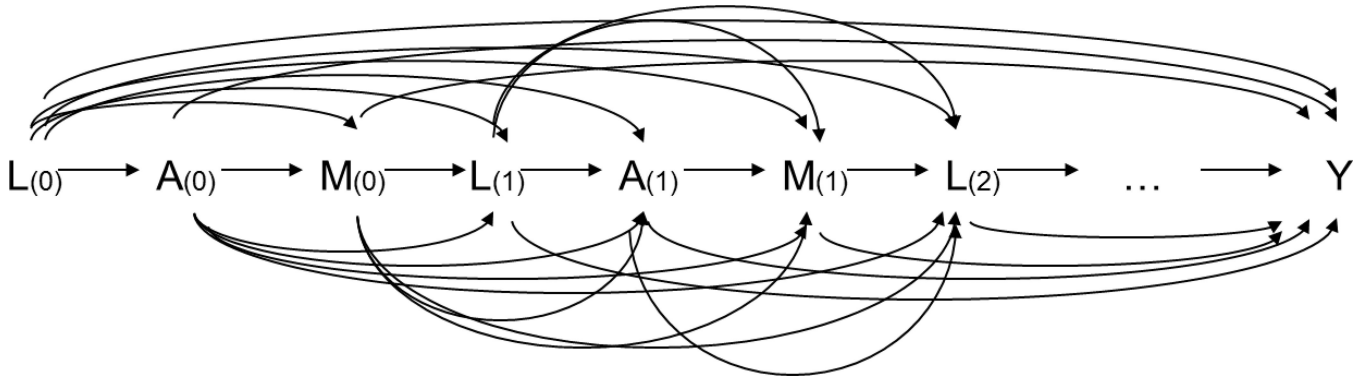

Figure 3.
Time-varying mediation with ordering of variables of A(t), M(t), L(t), for t = 0 to T−1.

Table 1

Summary of covariate models.

Variable	Type of model when used as dependent variable	Functional form when used as predictor
Non-modifiable		
Gender	Not predicted	Indicator
Age	Not predicted	Quadratic linear
Height	Not predicted	Quadratic linear
Education level	Not predicted	Six categories[a]
Occupation	Not predicted	Six categories[a]
Marital status	Not predicted	Three categories[b]
Baseline smoking	Not predicted	Three categories[b]
Modifiable		
SBP	Linear	Quadratic linear
Smoking	Logistic then log-linear[c]	Quadratic linear
Cholesterol	Linear	Quadratic linear
Anti-hypertension drug	Linear	Three categories[b]

[a] Education level categories are ≤ 8th grade, some high school, high school graduate, some college, college graduate, and post-graduate. Occupation categories are executive/supervisory, technical, laborer, clerical, sales, and housewife.

[b] Marital status categories are single, married, and divorce or widowed. Baseline smoking are smoking, not smoking, and quitting. Anti-hypertension drug are regular use, not use, and sporadic use.

[c] zero-continuous variables such as cigarettes per day are predicted in two stages, first a logistic regression on an indicator of whether the variable is nonzero and then a linear regression of the log of the nonzero values.

Table 2

Baseline characteristics of eligible participants grouped by former smoking status.

Characteristic	Quitters (n = 183)	Non-smokers (n = 1,174)	Smokers (n = 1,759)
Male (%)	128 (69)	188 (16)	1049 (59)
Age, year (Mean (SD))	50 (8.5)	49 (8.5)	46 (8.0)
SBP, mmHg (Mean (SD))	128 (18)	132 (22)	126 (18)
BMI (Mean (SD))	26 (4)	26 (4)	25 (4)
Chol (Mean (SD))	234 (44)	232 (45)	228 (43)
Height (Mean (SD))	66 (3)	63 (3)	65 (3)
Education (%)			
<High school	65 (36)	507 (43)	675 (38)
High school	64 (35)	420 (36)	707 (40)
College or higher	51 (28)	227 (19)	350 (20)
Missing	3 (2)	20 (2)	27 (2)
Work (%)			
Supervisory	58 (32)	156 (13.3)	419 (23.8)
Technical	20 (10.9)	31 (2.6)	130 (7.4)
Laborer	48 (26.2)	238 (20.3)	517 (29.4)
Clerical	12 (6.6)	76 (6.5)	109 (6.2)
Sales	8 (4.4)	19 (1.6)	91 (5.2)
Housewife	36 (19.7)	647 (55.1)	478 (27.2)
Missing	1 (0.5)	7 (0.6)	15 (0.9)
Marital (%)			
Married	163 (89.1)	935 (80)	1580 (90)
Single	10 (6)	145 (12)	105 (6)
Divorced	10 (6)	94 (8)	74 (4)

SBP: systolic blood pressure; BMI: body mass index; Chol: cholesterol level;

No one use the anti-hypertensive drugs at the beginning.

Table 3

Estimates of the total effect of smoking 20 cigarettes per day for 10 years (compared with no smoking) on SBP and BMI.

Intervention	SBP (mm-Hg)	Change in SBP (95% CI)	BMI (kg/m^2)	Change in BMI (95% CI)
No intervention	136.2	0.5 (−0.1, 1.0)	25.8	−0.00 (−0.2, 0.1)
No smoking	135.7	Reference	25.9	Reference
20 cigarettes/day	136.9	1.2 (−1.0, 3.1)	25.7	−0.2 (−0.4, 0.0)

SBP: systolic blood pressure; BMI: body mass index; CI: confident interval.

Table 4

Randomly interventional analogue of total effect, of natural direct effect, and of natural indirect effect for the effect of smoking 20 cigarettes per day for 10 years (compared with no smoking) on SBP, mediated by BMI change over time.

	Estimate	95% CI
$E[Y_{0G0}]$	135.691	134.93, 137.11
$E[Y_{1G0}]$	137.211	135.76, 138.80
$E[Y_{0G1}]$	135.336	134.57, 136.69
$E[Y_{1G1}]$	136.874	135.64, 138.37
rTE	1.18	−0.68, 2.69
rNDE	1.52	−0.25, 2.90
rNIE	−0.34	−0.52, −0.13

rTE: randomly interventional analogue of total effect; rNDE: randomly interventional analogue of natural direct effect; rNIE: randomly interventional analogue of natural indirect effect; CI: confident interval; E[Y0G0], E[Y1G0], E[Y0G1], and E[Y1G1]

represent “no smoking with BMI distributed as the BMI under no smoking”,
“smoking 20 cigarettes per day with BMI distributed as the BMI under no smoking”,
“no smoking with BMI distributed as the BMI under smoking 20 cigarettes per day”,
and “smoking 20 cigarettes per day with BMI distributed as the BMI under smoking 20 cigarettes per day”, respectively.