

# UCSF

## UC San Francisco Previously Published Works

### Title

Causal mediation analysis with multiple causally non-ordered mediators

### Permalink

<https://escholarship.org/uc/item/13p85868>

### Journal

Statistical Methods in Medical Research, 27(1)

### ISSN

0962-2802

### Authors

Taguri, Masataka  
Featherstone, John  
Cheng, Jing

### Publication Date

2018

### DOI

10.1177/0962280215615899

Peer reviewed



Published in final edited form as:

*Stat Methods Med Res.* 2018 January ; 27(1): 3–19. doi:10.1177/0962280215615899.

## Causal mediation analysis with multiple causally non-ordered mediators

Masataka Taguri<sup>1,2</sup>, John Featherstone<sup>2</sup>, and Jing Cheng<sup>2</sup>

<sup>1</sup>Department of Biostatistics, School of Medicine, Yokohama City University, Yokohama, Japan

<sup>2</sup>School of Dentistry, University of California, San Francisco, San Francisco, CA, USA

### Abstract

In many health studies, researchers are interested in estimating the treatment effects on the outcome around and through an intermediate variable. Such causal mediation analyses aim to understand the mechanisms that explain the treatment effect. Although multiple mediators are often involved in real studies, most of the literature considered mediation analyses with one mediator at a time. In this article, we consider mediation analyses when there are causally non-ordered multiple mediators. Even if the mediators do not affect each other, the sum of two indirect effects through the two mediators considered separately may diverge from the joint natural indirect effect when there are additive interactions between the effects of the two mediators on the outcome. Therefore, we derive an equation for the joint natural indirect effect based on the individual mediation effects and their interactive effect, which helps us understand how the mediation effect works through the two mediators and relative contributions of the mediators and their interaction. We also discuss an extension for three mediators. The proposed method is illustrated using data from a randomized trial on the prevention of dental caries.

### Keywords

Causal inference; effect decomposition; mediation analysis; multiple mediators; natural direct effect; natural indirect effect

## 1 Introduction

In many health studies, researchers are interested in estimating the treatment effects on the outcome around and through an intermediate variable (called a mediator), where the corresponding effects are called direct and indirect effects, respectively, and their sum is the total effect of the treatment on the outcome of interest. Such mediation analyses aim to understand the mechanisms that explain the treatment effect. Robins and Greenland<sup>1</sup> originally put forward a formal study of causal mediation analysis. Following their work,

Reprints and permissions: [sagepub.co.uk/journalsPermissions.nav](http://sagepub.co.uk/journalsPermissions.nav)

Corresponding author: Masataka Taguri, Yokohama City University, 3-9 Fukuura, Kanazawa-ku, Yokohama 236-0004, Japan. taguri@yokohama-cu.ac.jp.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Pearl<sup>2</sup> showed that a total effect can always be broken down into a natural direct and indirect effects. There is growing literature on evaluating natural direct and indirect effects.<sup>3–11</sup> Although a treatment often affects the outcome through multiple mediators in real studies, most of the literature considered a mediation analysis with a single mediator only.

When multiple intermediate variables are involved in a study, Daniel et al.<sup>12</sup> summarized existing approaches as the following three types in the setting of two mediators,  $M_1$  and  $M_2$  (see Figure 1(a) and (b)): (1)  $M_2$  is the mediator of interest, and  $M_1$  is treated as a mediator-outcome confounder affected by treatment, leading to a two-way decomposition into an effect (indirect) through  $M_2$  and an effect (direct) not through  $M_2$ <sup>13–17</sup>; (2) path-specific effects are estimated, but not in such a way that their sum equals the total effect<sup>18,19</sup>; and (3) the multiple mediators do not causally affect one another,<sup>20–22</sup> that is, the arrow from  $M_1$  to  $M_2$  in Figure 1(a) is assumed absent (Figure 1(b)). Imai and Yamamoto<sup>23</sup> considered approaches for all the three conditions assuming a linear structural equation model for the outcome and mediator. Daniel et al.<sup>12</sup> considered the finest possible decomposition of the total effect when there are two causally ordered mediators and evaluated each path-specific effect. In addition, VanderWeele and Vansteelandt<sup>24</sup> considered the multiple mediators one at a time as joint mediators and defined the “joint” natural direct and indirect effects as an extension of the usual two-way decomposition of the total effect with regression-based approach and weighting approach.

In this article, we will focus on the setting (3), that is, there are causally non-ordered multiple mediators, and provide an analytic approach to express the direct and indirect effects independent of models. Multiple mediators can be non-causally ordered when the treatment has multiple components targeting on multiple non-causally related mediators. For example, a cavity prevention plan for high-risk patients often has an antibacterial component to reduce oral bacteria as well as fluoride therapy to strength the teeth, where the two mediators oral bacteria level and fluoride level are not causally related.<sup>25</sup> With two causally non-ordered mediators involved, there are three path-specific effects from treatment ( $A$ ) to outcome ( $Y$ ): the direct pathway ( $A \rightarrow Y$ ), the indirect pathway through  $M_1$  only ( $A \rightarrow M_1 \rightarrow Y$ ), and the indirect pathway through  $M_2$  only ( $A \rightarrow M_2 \rightarrow Y$ ).

We can then estimate the three corresponding path-specific effects by separate analyses. Although the individual path-specific effects allow us to understand how the treatment works through individual paths alone, it may not give us the whole picture of the treatment effects involving multiple mediators. For example, consider the pathway through  $M_1$  only. The indirect effect will be the indirect effect of the treatment through  $M_1$ , and the direct effect will thus be the effect through all other pathways including the effect through  $M_2$ . However, in a real study, researchers are often interested in how the treatment works through individual mediators ( $M_1$  and  $M_2$ ) at the same time, including their possible interaction if there is any. In such a case, a joint natural indirect effect through ( $M_1, M_2$ ) will be of interest. We might then think that the sum of the two natural indirect effects for  $M_1$  and  $M_2$  considered separately should equal the joint natural indirect effect through ( $M_1, M_2$ ). In fact, even if the two mediators do not affect each other, the sum of two indirect effects considered separately may diverge from the joint indirect effect when there are additive interactions between the effects through the two mediators on the outcome.<sup>24</sup> Therefore, in this article,

we aim to provide an analytical approach to express the total effect as a function of the indirect effects through two causally non-ordered mediators ( $M_1$ ,  $M_2$ ) and the direct effect around ( $M_1$ ,  $M_2$ ) independent of models. The derived expression will help us understand how the mediation effect works through the two mediators and the relative contributions of different components and their interaction. Assumptions for identification, model-based estimation, and extension to more than two mediators will also be discussed in the article.

The remainder of this article is organized as follows: In Section 2, we briefly review the direct and indirect effects in the single mediator setting. In Section 3, we present the identification assumptions and review two existing approaches in the presence of two causally non-ordered mediators. In Section 4, we present a novel three-way decomposition of the joint natural indirect effect. In Section 5, we give the identification formula for our estimands. In Section 6, we discuss extensions to the three mediators settings, and to the vector-valued mediators. In Section 7, we apply the proposed method to data from a randomized trial evaluated the effect of a combined antibacterial and fluoride therapy on the prevention of dental caries. Finally, in Section 8, we conclude with a discussion.

## 2 A brief review of single mediator case

We first briefly review the direct and indirect effects for a single mediator. See Imai et al.<sup>9</sup> for a more detailed explanation. Let  $Y$  denote an observed outcome for an individual,  $A$  denote a binary treatment or exposure (1: treatment or exposed, 0: control or non-exposed),  $C$  denote a set of confounding variables that may affect the treatment, mediator, and/or outcome, and  $M$  denote a single potential mediator that may be on the pathway from the treatment to the outcome (Figure 2). There may be other mediators as well but when focusing on only one mediator, the effect through other mediators would be included in the direct path from  $A$  to  $Y$  not through  $M$ .

To conduct a causal mediation analysis, we use the potential outcome framework.<sup>26,27</sup> Let  $Y(a)$  and  $M(a)$  denote the potential outcome and potential mediator, respectively, that would be observed if, possibly contrary to the fact,  $A$  were set to  $a$ . Likewise, let  $Y(a,m)$  denote the potential outcome that would be observed if, possibly contrary to the fact,  $A$  were set to  $a$  and  $M$  were set to  $m$ . We also make assumptions referred to as the consistency and composition assumptions.<sup>7</sup> The consistency assumption for  $(A, M)$  is that among the subgroup with the observed treatment  $A=a$  and the observed mediator  $M=m$ , the observed outcome  $Y$  is equal to  $Y(a,m)$ . The consistency assumption for the effect of the treatment on the mediator is that among the subgroup with the observed treatment  $A=a$ , the observed mediator  $M$  is equal to  $M(a)$ . The composition assumption is that  $Y(a)=Y(a, M(a))$ .

Robins and Greenland<sup>1</sup> and Pearl<sup>2</sup> considered the natural direct effect of treatment  $A$  on outcome  $Y$ ,  $\{Y(1, M(0)) - Y(0, M(0))\}$ . This natural direct effect compares the potential outcome under treatment and control given the mediator  $M$  at its natural level under control  $M(0)$ , so is also referred as the “pure direct effect.”<sup>1</sup> The natural indirect effect  $\{Y(1, M(1)) - Y(1, M(0))\}$  they considered compares the potential outcome that would be observed when the subject is treated and mediator is changed from  $M(0)$  to  $M(1)$ . This natural indirect effect is also referred as the “total indirect effect.”<sup>1</sup> The total effect can then be decomposed into

the natural direct and indirect effect as:  $Y(1) - Y(0) = Y(1, M(1)) - Y(0, M(0)) = \{Y(1, M(1)) - Y(1, M(0))\} + \{Y(1, M(0)) - Y(0, M(0))\}$ . Alternatively, we can also decompose the total effect as:  $Y(1) - Y(0) = \{Y(1, M(1)) - Y(0, M(1))\} + \{Y(0, M(1)) - Y(0, M(0))\}$ , where  $\{Y(1, M(1)) - Y(0, M(1))\}$  is referred as the “total direct effect” and  $\{Y(0, M(1)) - Y(0, M(0))\}$  as the “pure indirect effect.”

Because we are not able to observe all the potential outcomes for one subject in a real study, the individual level effects cannot be identified. On the other hand, under some assumptions, the population average effects can be identified. Given confounders  $C=c$ , the population average effects are conditional expectations of the individual level effects  $E[Y(1) - Y(0)|c]$ ,  $E[Y(1, M(1)) - Y(1, M(0))|c]$ , and  $E[Y(1, M(0)) - Y(0, M(0))|c]$ . Various assumptions have been proposed for the identification of the population average natural direct and indirect effects. Most literature first assume no unmeasured confounding on three relationships.

**A1** *No-unmeasured confounding of the A-Y relation.*

$$Y(a, m) \perp\!\!\!\perp A|C \text{ for all } (a, m).$$

**A2** *No-unmeasured confounding of the M-Y relation.*

$$Y(a, m) \perp\!\!\!\perp M(a)|A=a, C \text{ for all } (a, m).$$

**A3** *No-unmeasured confounding of the A-M relation.*

$$M(a) \perp\!\!\!\perp A|C \text{ for all } a.$$

In addition, Pearl<sup>2</sup> made the following assumption for identification:

**A4** *A cross-world independence assumption.*

$$Y(a, m) \perp\!\!\!\perp M(a^*)|C \text{ for all } (a, a^*, m).$$

If we assume that data are generated from Pearl’s nonparametric structural equation model (NPSEM),<sup>28</sup> then A4 will hold if there is no mediator-outcome confounder that is affected itself by the treatment. If a mediator-outcome confounder is affected by the treatment, then without additional assumptions, natural direct and indirect effects cannot be nonparametrically identified even under Pearl’s NPSEM irrespective of such a confounder is measured or not.<sup>18</sup> Figure 2 shows a causal diagram that is compatible with assumptions A1–A4 under Pearl’s NPSEM.

### 3 Natural direct and indirect effects for two causally non-ordered mediators

#### 3.1 Notation and assumptions

We now consider the situation that there are two causally non-ordered mediators  $M_1$  and  $M_2$ , meaning that the causal relationship between  $M_1$  and  $M_2$  is absent as in Figure 1(b). In contrast, Figure 1(a) corresponds to a case where  $M_2$  is causally affected by  $M_1$ . In this

article, we will focus on the case illustrated in Figure 1(b). Let  $M_1(a)$ ,  $M_2(a)$ , and  $Y(a, m_1, m_2)$  be obvious extensions of the potential outcomes defined in Section 2. We also assume the consistency and composition assumptions for these potential outcomes. The observed outcome  $Y$  is equal to  $Y(A, M_1(A), M_2(A))$ . We assume that the potential mediator  $M_2(a, m_1) = M_2(a)$  does not depend on the value of  $m_1$ , implying that  $M_2$  is not causally affected by  $M_1$ .

We extend Assumptions A1–A4 to B1–B4 for two causally non-ordered mediators  $M_1$  and  $M_2$ .

**B1** *No-unmeasured confounding of the A–Y relation.*

$$Y(a, m_1, m_2) \perp\!\!\!\perp A | C \text{ for all } (a, m_1, m_2).$$

**B2** *No-unmeasured confounding of the ( $M_1, M_2$ )-Y relation.*

$$Y(a, m_1, m_2) \perp\!\!\!\perp \{M_1(a), M_2(a)\} | A=a, C \text{ for all } (a, m_1, m_2).$$

**B3** *No-unmeasured confounding of the A-( $M_1, M_2$ ) relation.*

$$M_k(a) \perp\!\!\!\perp A | C \text{ for all } (a, k).$$

**B4** *An extended cross-world independence assumption.*

$$Y(a, m_1, m_2) \perp\!\!\!\perp \{M_1(a^*), M_2(a^{**})\} | C, \\ M_1(a^*) \perp\!\!\!\perp M_2(a^{**}) | C, \text{ for all } (a, a^*, a^{**}, m_1, m_2).$$

Again, under the NPSEM, B4 will hold if there is no mediator-outcome confounder that is affected by the treatment. See Robins and Richardson<sup>29</sup> for a more detailed discussion on the NPSEM and its relation to other graphical causal models. Assumptions B1–B4 are sufficient to identify  $E[Y(a, M_1(a^*), M_2(a^{**})) | c]$  for all  $(a, a^*, a^{**})$ , as shown later in Section 5. In the remaining two subsections, we will review two existing approaches for mediation analysis in this context, and point out some counter-intuitive results.

### 3.2 A two-way decomposition of the total effect into the joint natural direct and indirect effects

Under the causal relationships in Figure 1(b), one may consider  $M_1$  and  $M_2$  as a joint mediator.<sup>24</sup> According to VanderWeele and Vansteelandt,<sup>24</sup> the natural direct and indirect effects with  $(M_1, M_2)$  as the mediator is defined by  $\{Y(1, M_1(0), M_2(0)) - Y(0, M_1(0), M_2(0))\}$  and  $\{Y(1, M_1(1), M_2(1)) - Y(1, M_1(0), M_2(0))\}$ , respectively. The joint natural indirect effect here is the treatment effect mediated through  $M_1$  or  $M_2$ , and the joint natural direct effect is the effect through neither  $M_1$  nor  $M_2$ . Then, the total effect is decomposed into the joint natural direct and indirect effects as follows:

$$Y(1) - Y(0) = \{Y(1, M_1(1), M_2(1)) - Y(1, M_1(0), M_2(0))\} + \{Y(1, M_1(0), M_2(0)) - Y(0, M_1(0), M_2(0))\} \\ = \text{total natural indirect effect} + \text{pure natural direct effect.}$$

(1)

The definitions given in (1) are natural extensions of the decomposition of the total effect into the total indirect effect and the pure direct effect to the two mediators setting. Another similar decomposition of the total effect into the joint total direct effect and the joint pure indirect effect is given as:  $Y(1) - Y(0) = \{Y(1, M_1(1), M_2(1)) - Y(0, M_1(1), M_2(1))\} + \{Y(0, M_1(1), M_2(1)) - Y(0, M_1(0), M_2(0))\} = \text{total natural direct effect} + \text{pure natural indirect effect.}$

The differences between the “pure” and “total” direct (indirect) effects are due to the differential inclusion of the interaction between the treatment and the mediators. In a single mediator case, VanderWeele<sup>30</sup> showed that the difference {total natural direct effect – pure natural direct effect} = {total natural indirect effect – pure natural indirect effect} corresponds to a “mediated interaction” between  $A$  and  $M$ , which is the product of an additive interaction of the treatment and the mediator on the outcome,  $\{Y(1,1) - Y(1,0) - Y(0,1) + Y(0,0)\}$ , and the effect of the treatment on the mediator,  $\{M(1) - M(0)\}$  (see Section 2 for the notation). This mediated interaction is arguably part of the effect that is mediated in the sense that it requires that the treatment changes the mediator.<sup>31</sup> In addition, under certain assumptions, the total indirect effect, in contrast to the pure indirect effect, would give more evidence for the actual operation of mediating mechanisms.<sup>32,33</sup> Thus, we will focus on the decomposition (1) in the remainder of this article for illustration. However, the methods discussed in the article should be directly applied to the other decomposition.

### 3.3 Two three-way decompositions of the joint natural indirect effect into path-specific natural indirect effects

If our aim is to compare the relative importance of  $M_1$  and  $M_2$  as a mediator, then we are interested in three path-specific effects from treatment to outcome: (i) the direct effect around the two mediators ( $A \rightarrow Y$ ), (ii) the indirect effect through  $M_1$  only ( $A \rightarrow M_1 \rightarrow Y$ ), and (iii) the indirect effects through  $M_2$  only ( $A \rightarrow M_2 \rightarrow Y$ ) (Figure 1(b)). Then, the joint natural indirect effect in (1) can be further decomposed into two path-specific effects as follows:

$$Y(1, M_1(1), M_2(1)) \\ - Y(1, M_1(0), M_2(0)) \\ = \{Y(1, M_1(1), M_2(1)) - Y(1, M_1(0), M_2(1))\} + \{Y(1, M_1(0), M_2(1)) \\ - Y(1, M_1(0), M_2(0))\} \quad (2)$$

$$= \{Y(1, M_1(1), M_2(0)) - Y(1, M_1(0), M_2(0))\} + \{Y(1, M_1(1), M_2(1)) - Y(1, M_1(1), M_2(0))\},$$

$$(3)$$

where the first terms in (2) and (3) are indirect effects through  $M_1$ , whereas the second terms in (2) and (3) are indirect effects through  $M_2$ . Daniel et al.<sup>12</sup> showed that there are six decompositions of the total effect into three path-specific effects. Of these six decompositions, Lange et al.<sup>22</sup> focused on (2) and (3) in conjunction with (1), whereas Imai and Yamamoto<sup>23</sup> considered other two decompositions. In this article, we will focus on (2) and (3) because these are only two decompositions such that the sum of the indirect effects through  $M_1$  and through  $M_2$  is equal to the joint total natural indirect effect in (1). For the notational convenience, we use  $PSE_1(a) = Y(1, M_1(1), M_2(a)) - Y(1, M_1(0), M_2(a))$  ( $a=0,1$ ) to denote indirect effects through  $M_1$ . Likewise, we use  $PSE_2(a) = Y(1, M_1(a), M_2(1)) - Y(1, M_1(a), M_2(0))$  ( $a=0,1$ ) to denote indirect effects through  $M_2$ . Using this notation, (2) =  $PSE_1(1) + PSE_2(0)$  and (3) =  $PSE_1(0) + PSE_2(1)$ .

Note that there would be no clear reason on which decomposition is preferred between (2) and (3), if we are interested in both  $M_1$  and  $M_2$ . However, the decompositions (2) and (3) will not necessarily give the same results when  $PSE_k(1) \neq PSE_k(0)$  ( $k=1,2$ ). If the analysis results from (2) and (3) diverge in the sense that indirect effects for  $M_1$  ( $M_2$ ) are different between these two decompositions, then there is no clear guidance on which decomposition to use. Note also that using obvious notation, the total natural indirect effect through  $M_1$  only,  $\{Y(1, M_1(1)) - Y(1, M_1(0))\}$ , can be written as<sup>24</sup>:

$$Y(1, M_1(1)) - Y(1, M_1(0)) = Y(1, M_1(1), M_2(1)) - Y(1, M_1(0), M_2(1)) = PSE_1(1), \quad (4)$$

which is equal to the first term of (2). Likewise, the total natural indirect effect through  $M_2$  only,  $\{Y(1, M_2(1)) - Y(1, M_2(0))\}$ , can be written as:

$$Y(1, M_2(1)) - Y(1, M_2(0)) = Y(1, M_1(1), M_2(1)) - Y(1, M_1(1), M_2(0)) = PSE_2(1), \quad (5)$$

which is equal to the second term of (3). From (2)–(5), we can understand that the sum of the two total natural indirect effect through  $M_1$  ( $=PSE_1(1)$ ) and  $M_2$  ( $=PSE_2(1)$ ) considered separately is not equal to the joint total natural indirect effect in general. As discussed in VanderWeele and Vansteelandt<sup>24</sup> and in Section 4, the sum of two indirect effects separately may diverge from the joint natural indirect effect when there are additive interactions between the two mediators on the outcome. Note that such interaction can arise even if the mediators do not affect each other.



## 4 A three-way decomposition of the joint natural indirect into path-specific natural indirect effects and an interactive effect

In Section 3, we have reviewed the existing approaches and discussed some potential problems. In this section, we present a new three-way decomposition of the joint natural indirect effect (and thus a four-way decomposition of the total effect) to resolve these problems. Here, we consider the setting of two binary mediators. Similar results are obtained for non-binary mediators (see Appendix 2). In Appendix 1, we show that the joint natural indirect effect can be further decomposed into the following three components:

$$\begin{aligned}
 & Y(1, M_1(1), M_2(1)) - Y(1, M_1(0), M_2(0)) \\
 = & \{Y(1, M_1(1), M_2(0)) - Y(1, M_1(0), M_2(0))\} + \{Y(1, M_1(0), M_2(1)) - Y(1, M_1(0), M_2(0))\} \\
 & + \{Y(1, M_1(1), M_2(1)) - Y(1, M_1(1), M_2(0)) - Y(1, M_1(0), M_2(1)) + Y(1, M_1(0), M_2(0))\} \\
 = & \{Y(1, M_1(1), M_2(0)) - Y(1, M_1(0), M_2(0))\} + \{Y(1, M_1(0), M_2(1)) - Y(1, M_1(0), M_2(0))\} \\
 & + \{Y(1, 1, 1) - Y(1, 1, 0) - Y(1, 0, 1) + Y(1, 0, 0)\} \{M_1(1) - M_1(0)\} \{M_2(1) - M_2(0)\}. \\
 = & \text{PSE}_1(0) + \text{PSE}_2(0) + \text{MI}
 \end{aligned}$$

(6)

The first component in the decomposition (6) is the indirect effect through  $M_1$  when the other mediator  $M_2$  is set to the control level, that is,  $\text{PSE}_1(0)$ . Likewise, the second component in the decomposition (6) is the indirect effect through  $M_2$  under the other mediator  $M_1$  is set to the control level, that is,  $\text{PSE}_2(0)$ . The third component in (6) is the product of the additive interaction between  $M_1$  and  $M_2$  with  $A=1$ ,  $\{Y(1, 1, 1) - Y(1, 1, 0) - Y(1, 0, 1) + Y(1, 0, 0)\}$ , the effect of the treatment on  $M_1$ ,  $\{M_1(1) - M_1(0)\}$ , and the effect of the treatment on  $M_2$ ,  $\{M_2(1) - M_2(0)\}$ . This interactive effect is nonzero if and only if the treatment affects both the two mediators and the additive interaction between  $M_1$  and  $M_2$  on  $Y$  is nonzero. Following the terminology of VanderWeele,<sup>30</sup> who considered a similar decomposition of natural direct and indirect effects in the case of a single mediator, we refer to this interactive effect as a “mediated interactive effect” or “mediated interaction” (MI) between  $M_1$  and  $M_2$ . This three-way decomposition includes the mediated interactive effect, so that it can be explicitly evaluated in a study and also resolves the ambiguity concerning the choice between (2) and (3). By definition, it follows that  $\text{MI} = Y(1, M_1(1), M_2(1)) - Y(1, M_1(1), M_2(0)) - Y(1, M_1(0), M_2(1)) + Y(1, M_1(0), M_2(0)) = \text{PSE}_1(1) - \text{PSE}_1(0) = \text{PSE}_2(1) - \text{PSE}_2(0)$ . Using these equalities, we obtain the following relations:  $\text{PSE}_1(1) = \text{PSE}_1(0) + \text{MI}$  and  $\text{PSE}_2(1) = \text{PSE}_2(0) + \text{MI}$ . Thus, we can understand that the difference between (2) and (3) are the differential inclusion of the mediated interaction for the indirect effect of  $M_1$  (decomposition (2)) or for the indirect effect of  $M_2$  (decomposition (3)). Thus, the results from (2) and (3) may diverge when there exists large additive interaction between the two mediators. Furthermore, using decomposition (6), we can understand how much of the joint natural indirect effect is explained by the interactive effect of the mediators as well as by each separate indirect effect.

Note that if the mediated interactive effect is equal to zero, then  $PSE_1(1)=PSE_1(0)$  and  $PSE_2(1)=PSE_2(0)$  hold. We then have that the joint total natural indirect effect in (6)  $=PSE_1(0)+PSE_2(0)=PSE_1(1)+PSE_2(1)$ , the same as the sum of the two separate total indirect effects, (4) and (5). Conversely, if the mediated interactive effect is nonzero, then (6) diverges from the sum of (4) and (5). This fact was noted by VanderWeele and Vansteelandt<sup>24</sup> by a similar but slightly different argument.

Note that we can consider another three-way decomposition of the joint natural indirect effect in (6) using  $PSE_1(1)$  and  $PSE_2(1)$  as follows:  $(6)=PSE_1(1)+PSE_2(1) -MI$ . In this article, we focus on the decomposition in (6) for illustration. However, the methods can be directly applied to the alternative decomposition. Here, we prefer the decomposition (6) for the following two reasons. First, these two three-way decompositions give essentially the same information considering the relations  $PSE_k(1)=PSE_k(0)+MI$  ( $k=1,2$ ) and it will be more natural and easier to interpret the result as the sum of  $PSE_1(0)$ ,  $PSE_2(0)$ , and  $MI$  rather than the difference between the sum of  $PSE_1(1)$ ,  $PSE_2(2)$ , and  $MI$ . Second, substantially,  $PSE_k(0)$  can be approximately interpreted as the effect if we had intervened only one mediator  $M_k$  while the other mediator held fixed at the control level. By contrast,  $PSE_k(1)$  can be approximately interpreted as the effect if we had intervened one mediator  $M_k$  with the other mediator held fixed at the intervention level. In our dental example in Section 7,  $PSE_1(0)$  would be interpreted as the indirect effect through  $M_1$  (bacteria level) if we had only implemented the antibacterial treatment (although the actual intervention was the combined antibacterial and fluoride therapy). On the other hand,  $PSE_1(1)$  would be interpreted as the indirect effect through  $M_1$  if we had implemented the antibacterial treatment for the population with the fluoride therapy uniformly implemented. Then,  $PSE_1(0)$  will be the more interesting quantity for understanding the effect through antibacterial therapy only. Similarly, we are also interested in the effect through fluoride therapy only when antibacterial therapy was not offered, that is,  $PSE_2(0)$ . Therefore, equation (6) provides a natural decomposition as the sum of effect through antibacterial therapy, effect through fluoride therapy, and their interactive effect.

Given the individual level decomposition (6), we can obtain a similar decomposition in the population average effect conditional on  $C=c$  by using B4 as follows:

$$\begin{aligned}
 & E[Y(1, M_1(1), M_2(1)) - Y(1, M_1(0), M_2(0)) | c] \\
 &= E[Y(1, M_1(1), M_2(0)) - Y(1, M_1(0), M_2(0)) | c] \\
 & \quad + E[Y(1, M_1(0), M_2(1)) - Y(1, M_1(0), M_2(0)) | c] \\
 & + E[Y(1, 1, 1) - Y(1, 1, 0) - Y(1, 0, 1) + Y(1, 0, 0) | c] E[M_1(1) - M_2(0) | c] E[M_2(1) - M_2(0) | c].
 \end{aligned}$$

(7)

In Appendix 2, we show the general formula which can be used for any type (i.e., non-binary) of mediators.

### 5 Identification

Under B1–B4, we obtain the following identification formula of  $E[Y(a, M_1(a^*), M_2(a^{**}))|c]$  for all  $(a, a^*, a^{**})$ :

$$E[Y(a, M_1(a^*), M_2(a^{**}))|c] = \sum_{m_1} \sum_{m_2} E[Y|a, m_1, m_2, c] p(m_1|a^*, c) p(m_2|a^{**}, c). \tag{8}$$

For continuous mediators, we simply replace sums by integrals in (8). Note that (8) is a special case of Theorem 1 in Daniel et al.,<sup>12</sup> but we show the derivation in Appendix 3 for completeness. In addition, under B1 and B2, we have  $E[Y(a, m_1, m_2)|c] = E[Y|a, m_1, m_2, c]$ , and  $\Pr(M_k(a) = m_k | c) = p(m_k|a, c)$  for  $k=(1,2)$ . Then, the all components in (6) as well as the joint natural direct effect can be identified from the observed data by the formulas given below:

$$\begin{aligned} & E[Y(1, M_1(0), M_2(0)) - Y(0, M_1(0), M_2(0))|c] \\ &= \sum_{m_1} \sum_{m_2} \{E[Y|A=1, m_1, m_2, c] - E[Y|A=0, m_1, m_2, c]\} p(m_1|A=0, c) p(m_2|A=0, c), \\ &= \sum_{m_1} \sum_{m_2} E[Y|A=1, m_1, m_2, c] \{p(m_1|A=1, c) - p(m_1|A=0, c)\} p(m_2|A=0, c), \\ &= \sum_{m_1} \sum_{m_2} E[Y|A=1, m_1, m_2, c] p(m_1|A=0, c) \{p(m_2|A=1, c) - p(m_2|A=0, c)\}, \\ &= \{E[Y(1, 1, 1) - Y(1, 1, 0) - Y(1, 0, 1) + Y(1, 0, 0)|c] E[M_1(1) - M_2(0)|c] E[M_2(1) - M_2(0)|c] \\ &+ E[Y|A=1, M_1=1, M_2=1|c] - E[Y|A=1, M_1=1, M_2=0|c] - E[Y|A=1, M_1=0, M_2=1|c] \\ &+ E[Y|A=1, M_1=0, M_2=0|c]\} \times \{E[M_1|A=1, c] - E[M_1|A=0, c]\} \\ &\quad \times \{E[M_2|A=1, c] - E[M_2|A=0, c]\}. \end{aligned}$$

(9)

See Appendix 2 for the general identification formula of the mediated interactive effect for a non-binary mediator.

With counterfactual average potential outcome (8) identified, we then consider a regression approach for the estimation of observables. Suppose that  $Y, M_1,$  and  $M_2$  are all continuous, and the following linear regression models hold for  $k=(1,2)$ :

$$\begin{aligned} E[Y|a, m_1, m_2, c] &= \theta_0 + \theta_1 a + \sum_{k=1}^2 \theta_2^{(k)} m_k + \sum_{k=1}^2 \theta_3^{(k)} a m_k + \theta_4 m_1 m_2 + \theta_5^T c, \\ E[M_k|a, c] &= \beta_0^{(k)} + \beta_1^{(k)} a + \beta_2^{(k)T} c. \end{aligned}$$

Then, by (8), we can show that for treatment levels  $a$  and  $a^*$ , the following equalities hold:

$$\begin{aligned}
 E[Y(a, M_1(a^*), M_2(a^*)) - Y(a^*, M_1(a^*), M_2(a^*)) | c] &= \{\theta_1 + \sum_{k=1}^2 \theta_3^{(k)} (\beta_0^{(k)} + \beta_1^{(k)} a^* + \beta_2^{(k)T} c)\} (a - a^*), \\
 E[Y(a, M_1(a), M_2(a^*)) - Y(a, M_1(a^*), M_2(a^*)) | c] &= \{(\theta_2^{(1)} + \theta_3^{(1)} a) \beta_1^{(1)} + \theta_4 \beta_1^{(1)} (\beta_0^{(2)} + \beta_1^{(2)} a^* + \beta_2^{(2)T} c)\} (a - a^*), \\
 E[Y(a, M_1(a^*), M_2(a)) - Y(a, M_1(a^*), M_2(a^*)) | c] &= \{(\theta_2^{(2)} + \theta_3^{(2)} a) \beta_1^{(2)} + \theta_4 \beta_1^{(2)} (\beta_0^{(1)} + \beta_1^{(1)} a^* + \beta_2^{(1)T} c)\} (a - a^*), \\
 E[Y(a, M_1(a), M_2(a)) - Y(a, M_1(a), M_2(a^*)) - Y(a, M_1(a^*), M_2(a)) + Y(a, M_1(a^*), M_2(a^*)) | c] &= \theta_4 \beta_1^{(1)} \beta_1^{(2)} (a - a^*) (a - a^*).
 \end{aligned}$$

(10)

From (10), we can see that the mediated interactive effect is equal to zero if the additive interaction between  $M_1$  and  $M_2$ , that is,  $\theta_4$ , is equal to zero. We can also see that if  $\theta_4$  is equal to zero, then the indirect effects through  $M_1$  and  $M_2$  are  $(\theta_2^{(1)} + \theta_3^{(1)} a) \beta_1^{(1)} (a - a^*)$  and  $(\theta_2^{(2)} + \theta_3^{(2)} a) \beta_1^{(2)} (a - a^*)$ , respectively. These are consistent with the results for the total indirect effect derived by VanderWeele and Vansteelandt<sup>7</sup> in a single mediator case. Standard errors of the estimators for (10) can be obtained using the delta method.

So far we considered effects conditional on the level of the covariates  $C=c$ . To obtain marginal effect estimates, we average the expressions in (8) over the marginal distribution of  $C$ . If at least one of the mediators is continuous and a linear regression model for  $Y$  does not hold, then we cannot generally obtain analytical formulas of (8) because we have to evaluate the integral on mediators. In such a case, we can use a Monte Carlo approach according to the method described in Imai et al.<sup>8</sup> to obtain marginal effect estimates. See Section 7 for the details in the application to the dental study. Standard errors and confidence intervals can be obtained based on the nonparametric bootstrap.

Another possible approach for the estimation is the inverse probability weighting (IPW).<sup>22</sup> We can obtain an estimator of  $E[Y(a, M_1(a^*), M_2(a^{**}))]$  by taking a weighted average of the outcome  $Y$  with the following weight  $w_i$  for the individual  $i$ :

$$w_i = \frac{I(A_i=a)}{\Pr(A_i=a|C_i=c_i)} \times \frac{\Pr(M_{1i}=m_{1i}|A_i=a^*, C_i=c_i)}{\Pr(M_{1i}=m_{1i}|A_i=a, C_i=c_i)} \times \frac{\Pr(M_{2i}=m_{2i}|A_i=a^{**}, C_i=c_i)}{\Pr(M_{2i}=m_{2i}|A_i=a, C_i=c_i)},$$

where  $I(\bullet)$  denotes the indicator function. In Appendix 4, we derive the asymptotic distribution of the IPW estimator. The regression approach above requires that the models for the outcome and mediators are correctly specified, whereas the weighting approach requires that the models for the treatment and mediators are correctly specified. The weighting approach will be more attractive if the mediation effects given in (10) is moderated by the confounders, in other words, if these effects depend on the value of  $c$ . This is because when we use the regression approach, it is difficult to specify the correct model for the outcome in the presence of the moderation by the confounders. However, the weighting approach will be unstable if some of the weights take very large values. When either the treatment or at least one of the mediators is continuous, then it is best not to use the weighting approach.

## 6 Extensions

### 6.1 Extension to the case of three mediators

Suppose that there are three causally non-ordered mediators (Figure 3(a)). Let  $Y(a, m_1, m_2, m_3)$  be a potential outcome if  $A, M_1, M_2, M_3$  were set to  $a, m_1, m_2, m_3$ , respectively. For notational convenience, define

$$\begin{aligned} \alpha_{12}(c) &= E[Y(1, 1, 1, 0) - Y(1, 0, 1, 0) - Y(1, 1, 0, 0) + Y(1, 0, 0, 0) | c], \\ \alpha_{13}(c) &= E[Y(1, 1, 0, 1) - Y(1, 0, 0, 1) - Y(1, 1, 0, 0) + Y(1, 0, 0, 0) | c], \\ \alpha_{23}(c) &= E[Y(1, 0, 1, 1) - Y(1, 0, 0, 1) - Y(1, 0, 1, 0) + Y(1, 0, 0, 0) | c], \\ \alpha_{123}(c) &= E[Y(1, 1, 1, 1) - Y(1, 1, 1, 0) - Y(1, 1, 0, 1) - Y(1, 0, 1, 1) + Y(1, 1, 0, 0) + Y(1, 0, 1, 0) + Y(1, 0, 0, 1) - Y(1, 0, 0, 0) | c], \end{aligned}$$

where  $\alpha_{12}(c)$ ,  $\alpha_{13}(c)$ , and  $\alpha_{23}(c)$  denote the two-way additive interaction between  $M_1$ - $M_2$ ,  $M_1$ - $M_3$ , and  $M_2$ - $M_3$ , respectively.  $\alpha_{123}(c)$  denotes the three-way additive interaction between  $M_1$ - $M_2$ - $M_3$ . At the individual level, the three-way additive interaction can be rewritten as follows:  $\{Y(1,1,1,1) - Y(1,1,0,1) - Y(1,0,1,1) + Y(1,0,0,1)\} - \{Y(1,1,1,0) - Y(1,1,0,0) - Y(1,0,1,0) + Y(1,0,0,0)\}$ . It will be nonzero if the two-way interaction between  $M_1$  and  $M_2$ ,  $\{Y(1,1,1,m_3) - Y(1,1,0,m_3) - Y(1,0,1,m_3) + Y(1,0,0,m_3)\}$ , is not constant across the level of  $m_3$ . Equivalently, it will be nonzero if the two-way interaction between  $M_1$  and  $M_3$ ,  $\{Y(1,1,m_2,1) - Y(1,1,m_2,0) - Y(1,0,m_2,1) + Y(1,0,m_2,0)\}$ , is not constant across the level of  $m_2$ , or if the two-way interaction between  $M_2$  and  $M_3$ ,  $\{Y(1,m_1,1,1) - Y(1,m_1,1,0) - Y(1,m_1,0,1) + Y(1,m_1,0,0)\}$ , is not constant across the level of  $m_1$ . Thus,  $\alpha_{123}(c)$  describes how the two-way interactions vary across the level of the other mediator.

Suppose that the NPSEM and the mechanism shown in Figure 3(a) hold for the full data including potential outcomes. Then, following a similar argument for the derivation of (6) and (7) (see Appendices 1 and 2), we obtain the following decomposition of the joint natural indirect effect:

$$\begin{aligned} & E[Y(1, M_1(1), M_2(1), M_3(1)) - Y(1, M_1(0), M_2(0), M_3(0)) | c] \\ &= E[Y(1, M_1(1), M_2(0), M_3(0)) - Y(1, M_1(0), M_2(0), M_3(0)) | c] \\ &+ E[Y(1, M_1(0), M_2(1), M_3(0)) - Y(1, M_1(0), M_2(0), M_3(0)) | c] \\ &+ E[Y(1, M_1(0), M_2(0), M_3(1)) - Y(1, M_1(0), M_2(0), M_3(0)) | c] \\ &+ \sum_{k=1}^2 \sum_{k' > k}^3 \delta_k(c) \delta_{k'}(c) \alpha_{kk'}(c) + \delta_1(c) \delta_2(c) \delta_3(c) \alpha_{123}(c), \end{aligned} \tag{11}$$

where  $\delta_k(c)$  ( $k=1,2,3$ ) denotes the effect of the treatment on  $M_k$ ,  $E[M_k(1) - M_k(0) | c]$ . We refer to the last term of (11) as the mediated interactive effect between  $M_1, M_2$ , and  $M_3$ . Note that the corresponding decomposition also holds at the individual level. From (11), we see that the joint natural indirect effect is decomposed into the three indirect effect through each mediator with the other mediators at the control level, three two-way mediated interactive effects, and one three-way mediated interactive effect. Note that the three-way mediated interactive effect at the individual level,  $\{M_1(1) - M_1(0)\} \{M_2(1) - M_2(0)\} \{M_3(1) - M_3(0)\} \{Y(1,1,1,1) - Y(1,1,1,0) - Y(1,1,0,1) + Y(1,0,1,1) + Y(1,1,0,0) + Y(1,0,1,0) +$

$Y(1,0,0,1) - Y(1,0,0,0)$ , is non-zero if and only if the treatment affects all three mediators, and the three-way additive interaction is non-zero.

### 6.2 Relaxing the identification assumptions

In previous sections, we have discussed mediation analysis when the mediators do not affect each other. In this section, we will consider the effects when two mediators lie on a causal pathway. Suppose that there are three mediators and that  $M_2$  causally affects  $M_3$  as in Figure 3(b). Then, the result described in Section 6.1 does not apply. Nevertheless, we consider  $M_2$  and  $M_3$  as a joint mediator,  $\mathbf{M}_2=(M_2, M_3)$ . Let  $\mathbf{M}_2(a)$  be the potential value of mediator  $\mathbf{M}_2$  that would be observed when the subject had treatment  $a$ , and let  $Y(a,m_1,\mathbf{m}_2)$  denote the potential outcome that would be observed when the subject had treatment  $a$  and mediators  $m_1$  and  $\mathbf{m}_2$ . We can define the joint natural indirect effect in a similar way as before but simply replace  $M_2$  with the vector of two mediators  $\mathbf{M}_2$ :  $\{Y(1,M_1(1), \mathbf{M}_2(1)) - Y(1, M_1(0), \mathbf{M}_2(0))\}$ . Then, the joint natural indirect effect can be decomposed into the indirect effect through  $M_1$ , the indirect effect through  $\mathbf{M}_2$ , and the mediated interactive effect between  $M_1$  and  $\mathbf{M}_2$ :  $\{Y(1, M_1(1), \mathbf{M}_2(1)) - Y(1, M_1(0), \mathbf{M}_2(0))\} = \{Y(1, M_1(1), \mathbf{M}_2(0)) - Y(1, M_1(0), \mathbf{M}_2(0))\} + \{Y(1, M_1(0), \mathbf{M}_2(1)) - Y(1, M_1(0), \mathbf{M}_2(0))\} + \sum_{m_1} \sum_{\mathbf{m}_2} Y(1,m_1, \mathbf{m}_2) \{I(M_1(1) = m_1) - I(M_1(0) = m_1)\} \{I(\mathbf{M}_2(1) = \mathbf{m}_2) - I(\mathbf{M}_2(0) = \mathbf{m}_2)\}$ . Note that we can easily see that the mediated interactive effect above is non-zero only if the treatment affects  $M_1$  and  $\mathbf{M}_2$ , and the additive interaction between  $M_1$  and  $\mathbf{M}_2$  is non-zero, that is,  $\{Y(1,1,\mathbf{m}_2) - Y(1,0,\mathbf{m}_2)\}$  is not always constant across the levels of  $\mathbf{m}_2$ . A sufficient condition for identification of  $E[Y(a, M_1(a^*), \mathbf{M}_2(a^{**}))|c]$  for all  $(a, a^*, a^{**})$  is similar to B1–B4 in Section 5, but now we need to replace  $M_2$  with  $\mathbf{M}_2$ . These conditions hold in Figure 3(a) to 3(c).

Note that  $M_2$  in Figure 3(b) may not be a mediator of interest but be treated as a confounder for the  $M_3$ -  $Y$  relation,<sup>16</sup> where  $M_2$  confounds the  $M_3$ -  $Y$  relation and, at the same time, it is affected by the treatment  $A$ . In this case, B4 is violated even if  $M_2$  is included in the confounder set  $C$  so that the results derived in Section 4 cannot be used if  $M_1$  and  $M_3$  are the two mediators we are interested in. However, we can consider  $\mathbf{M}_2=(M_2, M_3)$  and then the result described in this section can still apply.

Moreover, consider a situation that there is an unmeasured common cause  $U$  of  $M_2$  and  $M_3$  as in Figure 3(c). In this case, even if  $M_2$  does not affect  $M_3$  (i.e., there is no arrow from  $M_2$  to  $M_3$ ), the estimates of effects through the mediators  $M_1$ ,  $M_2$ , and  $M_3$  separately will be biased because when  $M_2$  ( $M_3$ ) is considered alone,  $U$  will be an unmeasured confounder for the  $M_2$ -  $Y$  ( $M_3$ -  $Y$ ) relation. However, when  $M_2$  and  $M_3$  are considered jointly,  $U$  no longer serves as a confounder. See VanderWeele and Vansteelandt<sup>24</sup> for a related discussion on the merit of considering mediators jointly.

## 7 Application

In this section, we will use the methods developed in this article to understand the mediation effects in The Caries Management by Risk Assessment (CAMBRA) randomized controlled clinical trial.<sup>34</sup> CAMBRA was a randomized controlled trial, which aimed to assess whether combined antibacterial and fluoride therapy based on risk assessment has beneficial effects on preventing new caries over 24 months follow-up in adults with one to seven baseline

cavitated teeth which were treated before initiating preventive therapies: the antibacterial therapy aimed to reduce oral bacteria whereas the fluoride therapy aimed to increase the fluoride level to strengthen teeth. In the study, participants in the control group ( $A=0$ ) received conventional treatment per usual practices (e.g., oral hygiene instruction, periodic dental cleaning and oral examination scheduled every 6 month, radiographs scheduled every 24 month, and restorative treatment as needed), whereas participants in the intervention group ( $A=1$ ) received a combined antibacterial (0.12% chlorhexidine gluconate mouth rinse) and fluoride therapy (1100 ppm sodium fluoride toothpaste, 0.05% sodium fluoride mouth rinse, and topical 1.1% NaF gel application). The primary analyses showed that the intervention group had a statistically significantly lower caries risk at follow-up and suggested a lower average caries increment compared with control over 24 months.<sup>34</sup> Our interest in this mediation analysis is whether this overall intervention effect was due mainly to bacteria reduction through antibacterial therapy, fluoride increase through fluoride therapy, or both. If observed, mediation through these mechanisms would provide further evidence for focusing future caries prevention efforts on these components.

The potential mediators of interest are two salivary oral bacteria (mutans streptococci (MS) and lactobacilli (LB)) levels and salivary fluoride level at 12 months. To make our identification assumptions more plausible, we consider MS and LB levels as a vector of mediators ( $\mathbf{M}_1$ ) and consider fluoride level as the other mediator ( $M_2$ ), where  $\mathbf{M}_1$  and  $M_2$  are assumed work through independent pathways. The outcome of interest ( $Y$ ) was the increment from baseline in the number of decayed, missing, and filled permanent surfaces (DMFS) at 24 months. DMFS is a nonnegative integer count with higher values indicating worse dental outcomes. From a total of 231 participants randomized, 101 (intervention group: 51; control group: 50) patients who had completed data on DMFS and relevant covariates were analyzed in this article.

Table 1 shows participants' characteristics at baseline by intervention group. Variables that were included in the set of  $C$  were: age, sex (male/female), race (Asian/black/white/Hispanic and others), education (high school/college/graduate or professional), timing of last dental visit (less than 1 year/2 to 3 years/3 years or more), brushed 2 times or more yesterday (yes/no), used fluoride toothpaste (yes/no), fair or poor oral health (yes/no), drank alcohol in past week (yes/no), and smoked cigarette within 30 days (yes/no). Because of the positively skewed distributions, all of the three mediators were log-transformed before the mediation analysis.

To assess whether the mediators do not affect one another (or more precisely, whether the mediators are independent of one another conditional on intervention and covariates), we first modeled  $p(m_2 | \mathbf{m}_1, a, c)$  using a linear regression model including main effects of  $\mathbf{M}_1$ ,  $A$ , and  $C$ . The  $p$ -values for MS and LB levels were 0.345 and 0.542, respectively, indicating no evidence for the violation of the assumption in the data. We therefore used the result described in Section 6.2 to estimate the direct effect, indirect effect, and mediated interaction. We modeled  $p(y | a, \mathbf{m}_1, m_2, c)$  with a negative binomial regression and the conditional distributions of the three mediators with linear regression models assuming normally distributed errors. In addition to the main effects of all the covariates, we included interaction terms between  $A$  and  $\mathbf{M}_1$ ,  $A$  and  $M_2$ , and  $\mathbf{M}_1$  and  $M_2$ , in the model for the

outcome. According to Imai et al.,<sup>8</sup> we conducted the analyses using the following steps for the estimation of the marginal effect  $E[Y(a, \mathbf{M}_1(a^*), M_2(a^{**}))]$ : (1) Fit the mediators and outcome models with observed data as described above. (2) Draw  $\mathbf{m}_1^*$  and  $m_2^*$  from  $\hat{p}(\mathbf{m}_1 | a^*, c)$  and  $\hat{p}(m_2 | a^{**}, c)$ . (3) Draw  $y^*$  from  $\hat{p}(y | a, \mathbf{m}_1^*, m_2^*, c)$ . (4) Perform Monte Carlo replications by repeating (2) and (3) 1000 times each. (5) Compute the mean of  $y^*$  over all individuals and Monte Carlo replications. 95% confidence intervals (CIs) were constructed based on the nonparametric bootstrap with 1000 resamples.

Table 2 shows various estimated direct and indirect effects of the intervention on DMFS at 2 years around and through its effect on participants' salivary bacteria and fluoride levels at 12 months. The estimated joint natural direct effect was  $-0.298$  (95% CI:  $-1.894$  to  $1.805$ ), and the joint natural indirect effect was  $-0.490$  (95% CI:  $-1.652$  to  $0.172$ ), and thus the total effect was  $-0.298 + (-0.490) = -0.788$  (95% CI:  $-2.108$  to  $0.847$ ). Applying the proposed three-way decomposition of the joint natural indirect effect, the indirect effect through  $\mathbf{M}_1$  only was  $-0.373$  ( $-1.541$  to  $0.195$ ), the indirect effect through  $M_2$  only was  $-0.022$  (95% CI:  $-0.366$  to  $0.789$ ), and the mediated interactive effect of  $\mathbf{M}_1$  and  $M_2$  was  $-0.095$  (95% CI:  $-0.807$  to  $0.171$ ). Thus, of the total effect,  $-0.298/-0.788=37.8\%$  was attributable to the joint natural direct effect,  $-0.373/-0.788=47.3\%$  was attributable to the indirect effect through  $\mathbf{M}_1$  only,  $-0.022/-0.788=2.8\%$  was attributable to the indirect effect through  $M_2$  only, and  $-0.095/-0.788=12.1\%$  was attributable to the mediated interaction. The overall proportion mediated was  $47.3 + 2.8 + 12.1=62.2\%$ . The results indicate that the effect of the intervention ( $A$ ) on DMFS ( $Y$ ) was mainly through its effect in decreasing salivary oral bacteria levels ( $\mathbf{M}_1$ ), although the effect is not significant due to smaller sample size for this analysis compared with primary analysis. Of the mediated effect, only a small portion of the effect was due to the effect in increasing salivary fluoride levels ( $M_2$ ). However, the moderated size of the mediated interactive effect of  $\mathbf{M}_1$  and  $M_2$  (more than 10%) indicates the effect of increased salivary fluoride level on DMFS through its interaction with decreased oral bacterial level, which agree with the results by Cheng et al.,<sup>25</sup> where a single composite mediator "overall caries risk" was considered based on the joint values of salivary bacterial and fluoride levels with more participants included in the analysis.

## 8 Discussion

In this article, we consider the joint natural indirect effect between two mediators as a function of the indirect effect for each mediator and the mediated interactive effect under the assumption that the mediators are not causally ordered. This relation holds even in cases when there is nonzero mediated interaction and thus a joint natural indirect effect cannot be a simple sum of the two individual natural indirect effects through the two mediator considered separately. Compared with the existing approaches described in Section 3, our proposed decomposition provides us an additional insight into the importance of the interaction among the mediators on the mediation mechanism. Such an insight may help the researchers or policymakers make a decision for a better intervention. As an example, consider a situation that there are two mediators and their mediated interactive effect is large relative to the sum of two individual mediation effects under the reference condition of the other mediator. This additional knowledge can help researchers or policymakers understand



that an intervention that affects one mediator but not the other may not be the optimal option because an intervention works best when it affects both of these mediators.

We have focused on cases where the mediators do not affect each other in this article. This assumption may hold when the intervention consists of multiple independent components, which work together to improve the outcome but one component would not causally affect the other component. As in our example, in the dental study, the intervention has two components: a combined antibacterial and fluoride therapy, where the antibacterial treatment reduced oral bacteria whereas the fluoride treatment made the teeth strong. Then, the oral bacterial load and fluoride level measured at follow-up time point after the intervention would be causally non-ordered mediators between the intervention and dental outcomes, as the oral bacterial load mainly captures the effect of the antibacterial therapy, whereas the fluoride level captures the effect of the fluoride therapy.

Even if  $M_2$  is causally affected by  $M_1$  as Figure 1(a), our proposed decomposition of the joint natural indirect effect still holds. To be more precise, let  $M_2(a) = M_2(a, M_1(a))$  denote the potential outcome of  $M_2$  if  $A$  were set to  $a$ . Then, we obtain exactly the same expression as (6). However, (6) is not based on a finest possible decomposition in this case, because  $\{Y(1, M_1(0), M_2(1)) - Y(1, M_1(0), M_2(0))\} = \{Y(1, M_1(0), M_2(1, M_1(1))) - Y(1, M_1(0), M_2(0, M_1(0)))\}$  is the sum of the two path-specific effects  $A \rightarrow M_2 \rightarrow Y$ ,  $\{Y(1, M_1(0), M_2(1, M_1(1))) - Y(1, M_1(0), M_2(0, M_1(1)))\}$ , and  $A \rightarrow M_1 \rightarrow M_2 \rightarrow Y$ ,  $\{Y(1, M_1(0), M_2(0, M_1(1))) - Y(1, M_1(0), M_2(0, M_1(0)))\}$ . Furthermore,  $E[Y(a, M_1(a^*), M_2(a^{**})) | c]$  is not nonparametrically identified under Pearl's NPSEM and Figure 1(a) except for the special case  $a^* = a^{**}$ .<sup>12</sup> Thus, we must impose additional strong assumptions for identification, or need to conduct a sensitivity analysis as described in Daniel et al.<sup>12</sup> Further research is needed on how to extend our method to the case of causally ordered multiple mediators.

We assume that there are no unmeasured confounders between the treatment and outcome (Assumption B1), the mediators and outcome (Assumption B2), and the treatment and mediators (Assumption B3). Although Assumptions B1 and B3 usually hold in a randomized trial, Assumption B2 does not necessarily hold even under the randomization of the treatment because the mediator cannot be randomized in a real study. In our analysis, we accounted for many potential confounders that may affect both the mediators and outcome. Even so, there may be utility in conducting sensitivity analyses that examine the effect of violations of B2. In the single mediator setting, there are many works that propose a sensitivity analysis method,<sup>15,17,23,35,36</sup> including the derivation of bounds for the natural direct and indirect effects.<sup>4,5,11,17,37</sup> Investigation into how such methods can be adapted to this setting is another important area for future research.

## Acknowledgments

We thank a referee for his/her insightful comments that helped improve the article.

### Funding

This work was partially supported by the overseas training program of the Japanese Society of Clinical Pharmacology and Therapeutics, by Grant-in-Aid for Scientific Research (No. 15K15951) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, and by grants U54 DE 019285 from the National Institute of Dental and Craniofacial Research (NIDCR), a component of the National Institutes of Health, which is

part of the U.S. Department of Health and Human Services. The CAMBRA trial was completed with support from NIDCR grant R01DE012455.

## References

1. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992; 3:143–155. [PubMed: 1576220]
2. Pearl, J. Direct and indirect effects. Proceedings of the seventeenth conference on uncertainty in artificial intelligence; San Francisco, CA: Morgan Kaufmann; 2001. p. 411-420.
3. van der Laan MJ, Petersen ML. Direct effect models. *Int J Biostat*. 2008; 4:1–27.
4. Kaufman S, Kaufman JS, MacLehose RF. Analytic bounds on causal risk differences in directed acyclic graphs involving three observed binary variables. *J Stat Plan Inference*. 2009; 139:3473–3487. [PubMed: 20161106]
5. Sjölander A. Bounds on natural effects in the presence of confounded intermediate variables. *Stat Med*. 2009; 28:558–571. [PubMed: 19035530]
6. VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology*. 2009; 20:18–26. [PubMed: 19234398]
7. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Stat Interface*. 2009; 2:457–468.
8. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods*. 2010; 15:309–334. [PubMed: 20954780]
9. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Stat Sci*. 2010; 25:51–71.
10. Daniels MJ, Roy JA, Kim C, et al. Bayesian inference for the causal effect of mediation. *Biometrics*. 2012; 68:1028–1036. [PubMed: 23005030]
11. Chiba Y, Taguri M. Alternative monotonicity assumptions for improving bounds on natural direct effects. *Int J Biostat*. 2013; 9:235–249. [PubMed: 23893690]
12. Daniel RM, De Stavola BL, Cousens SN, et al. Causal mediation analysis with multiple mediators. *Biometrics*. 2015; 71:1–14. [PubMed: 25351114]
13. Vansteelandt S, VanderWeele TJ. Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions. *Biometrics*. 2012; 68:1019–1027. [PubMed: 22989075]
14. Tchetgen Tchetgen EJ, VanderWeele TJ. Identification of natural direct effects when a confounder of the mediator is directly affected by exposure. *Epidemiology*. 2014; 25:282–291. [PubMed: 24487211]
15. VanderWeele TJ, Chiba Y. Sensitivity analysis for direct and indirect effects in the presence of exposure-induced mediator-outcome confounders. *Epidemiol Biostat Public Health*. 2014; 11:e9027. [PubMed: 25580387]
16. VanderWeele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology*. 2014; 25:300–306. [PubMed: 24487213]
17. Taguri M, Chiba Y. A principal stratification approach for evaluating natural direct and indirect effects in the presence of treatment-induced intermediate confounding. *Stat Med*. 2015; 34:131–144. [PubMed: 25312003]
18. Avin, C., Shpitser, I., Pearl, J. Identifiability of path-specific effects. Proceedings of the international joint conference on artificial intelligence; Edinburgh, UK: Morgan-Kaufmann; 2005. p. 357-363.
19. Albert JM, Nelson S. Generalized causal mediation analysis. *Biometrics*. 2011; 67:1028–1038. [PubMed: 21306353]
20. MacKinnon, DP. Multivariate applications in substance use research. Mahwah, NJ: Lawrence Erlbaum Associates Publishers; 2000. Contrasts in multiple mediator models; p. 141-160.
21. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods*. 2008; 40:879–891. [PubMed: 18697684]
22. Lange T, Rasmussen M, Thygesen LC. Assessing natural direct and indirect effects through multiple pathways. *Am J Epidemiol*. 2014; 179:513–518. [PubMed: 24264291]

23. Imai K, Yamamoto T. Identification and sensitivity analysis for multiple causal mechanisms: revisiting evidence from framing experiments. *Political Anal.* 2013; 21:141–171.
24. VanderWeele TJ, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiol Methods.* 2013; 2:95–115.
25. Cheng J, Chaffee BW, Cheng NF, et al. Understanding treatment effect mechanisms of the CAMBRA randomized trial in reducing caries increment. *J Dental Res.* 2015; 94:44–51.
26. Neyman J. On the application of probability theory to agricultural experiments: essay on principles, Section 9. *Ann Agric Sci.* 1923 Translated in *Stat Sci* 1990; 5: 465–472.
27. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol.* 1974; 66:688–701.
28. Pearl, J. *Causality: models, reasoning, and inference.* 2. New York: Cambridge University Press; 2009.
29. Robins, JM., Richardson, TS. *Causality and psychopathology: finding the determinants of disorders and their cures.* New York, NY: Oxford University Press; 2010. Alternative graphical causal models and the identification of direct effects; p. 103-158.
30. VanderWeele TJ. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology.* 2013; 24:224–232. [PubMed: 23354283]
31. VanderWeele TJ. A unification of mediation and interaction: a four-way decomposition. *Epidemiology.* 2014; 25:749–761. [PubMed: 25000145]
32. VanderWeele TJ. Subtleties of explanatory language: what is meant by “mediation”? *Eur J Epidemiol.* 2011; 26:343–346. [PubMed: 21553297]
33. Suzuki E, Yamamoto E, Tsuda T. Identification of operating mediation and mechanism in the sufficient-component cause framework. *Eur J Epidemiol.* 2011; 26:347–357. [PubMed: 21448741]
34. Featherstone JD, White JM, Hoover CI, et al. A randomized clinical trial of anticaries therapies targeted according to risk assessment (caries management by risk assessment). *Caries Res.* 2012; 46:118–129. [PubMed: 22472515]
35. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology.* 2010; 21:540–551. [PubMed: 20479643]
36. Tchetgen Tchetgen EJ, Shpitser I. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Ann Stat.* 2012; 40:1816–1845. [PubMed: 26770002]
37. Tchetgen Tchetgen EJ, Phiri K. Bounds for pure direct effect. *Epidemiology.* 2014; 25:775–776. [PubMed: 25076155]
38. Tsiatis, AA. *Semiparametric theory and missing data.* New York: Springer; 2006.

### Appendix 1. Derivation of (6)

We first note that the difference between the joint natural indirect effect,  $\{Y(1, M_1(1), M_2(1)) - Y(1, M_1(0), M_2(0))\}$ , and sum of the indirect effects for  $M_1$ ,  $\{Y(1, M_1(1), M_2(0)) - Y(1, M_1(0), M_2(0))\}$ , and that for  $M_2$ ,  $\{Y(1, M_1(0), M_2(1)) - Y(1, M_1(0), M_2(0))\}$ , are given by:

$$Y(1, M_1(1), M_2(1)) - Y(1, M_1(1), M_2(0)) - Y(1, M_1(0), M_2(1)) + Y(1, M_1(0), M_2(0)). \tag{12}$$

For our objective, it is enough to show that (12) is equal to the third component of (6).

First, we consider to evaluate  $M_1(1)$  and  $M_1(0)$  under the fixed values of  $M_2(1)$  and  $M_2(0)$ . If  $M_1(1) = M_1(0)$ , then  $(12) = Y(1, M_1(1), M_2(1)) - Y(1, M_1(1), M_2(0)) - Y(1, M_1(1), M_2(1)) + Y(1, M_1(1), M_2(0)) = 0$ . If  $M_1(1) - M_1(0) = 1$ , that is,  $M_1(1) = 1$  and  $M_1(0) = 0$ , then  $(12) = Y(1, 1, M_2(1)) - Y(1, 1, M_2(0)) - Y(1, 0, M_2(1)) + Y(1, 0, M_2(0)) = \{Y(1, 1,$

$M_2(1) - Y(1,1, M_2(0)) - Y(1,0, M_2(1)) + Y(1,0, M_2(0))\} \{M_1(1) - M_1(0)\}$ . If  $M_1(1) - M_1(0) = -1$ , that is,  $M_1(1) = 0$  and  $M_1(0) = 1$ , then (12) =  $Y(1,0, M_2(1)) - Y(1,0, M_2(0)) - Y(1,1, M_2(1)) + Y(1,1, M_2(0)) = - \{ Y(1,1, M_2(1)) - Y(1,1, M_2(0)) - Y(1,0, M_2(1)) + Y(1,0, M_2(0))\} = \{ Y(1,1, M_2(1)) - Y(1,1, M_2(0)) - Y(1,0, M_2(1)) + Y(1,0, M_2(0))\} \{M_1(1) - M_1(0)\}$ . Thus, (12) can be expressed as follows:

$$\{Y(1, 1, M_2(1)) - Y(1, 1, M_2(0)) - Y(1, 0, M_2(1)) + Y(1, 0, M_2(0))\} \{M_1(1) - M_2(0)\}. \quad (13)$$

Next, we consider to evaluate  $M_2(1)$  and  $M_2(0)$ . If  $M_2(1) = M_2(0)$ , then (13) =  $\{ Y(1,1, M_2(1)) - Y(1,1, M_2(1)) - Y(1,0, M_2(1)) + Y(1,0, M_2(1))\} \{M_1(1) - M_1(0)\} = 0$ . If  $M_2(1) - M_2(0) = 1$ , that is,  $M_2(1) = 1$  and  $M_2(0) = 0$ , then (13) =  $\{ Y(1,1,1) - Y(1,1,0) - Y(1,0,1) + Y(1,0,0)\} \{M_1(1) - M_1(0)\} = \{ Y(1,1,1) - Y(1,1,0) - Y(1,0,1) + Y(1,0,0)\} \{M_1(1) - M_1(0)\} \{M_2(1) - M_2(0)\}$ . If  $M_2(1) - M_2(0) = -1$ , that is,  $M_2(1) = 0$  and  $M_2(0) = 1$ , then (13) =  $\{ Y(1,1,0) - Y(1,1,1) - Y(1,0,0) + Y(1,0,1)\} \{M_1(1) - M_1(0)\} = - \{ Y(1,1,1) - Y(1,1,0) - Y(1,0,1) + Y(1,0,0)\} \{M_1(1) - M_1(0)\} = \{ Y(1,1,1) - Y(1,1,0) - Y(1,0,1) + Y(1,0,0)\} \{M_1(1) - M_1(0)\} \{M_2(1) - M_2(0)\}$ . Thus, (12) is always equal to  $\{ Y(1,1,1) - Y(1,1,0) - Y(1,0,1) + Y(1,0,0)\} \{M_1(1) - M_1(0)\} \{M_2(1) - M_2(0)\}$ , as desired.

## Appendix 2. General expression of the mediated interactive effect for two mediators

We consider general types of treatments and mediators. We will compare two treatment levels,  $a$  and  $a^*$ . The mediated interactive effect between  $M_1$  and  $M_2$  at the individual level is defined as the difference between the joint natural indirect effect,  $\{ Y(a, M_1(a), M_2(a)) - Y(a, M_1(a^*), M_2(a^*))\}$ , and sum of the indirect effects for  $M_1$ ,  $\{ Y(a, M_1(a), M_2(a^*)) - Y(a, M_1(a^*), M_2(a^*))\}$ , and that for  $M_2$ ,  $\{ Y(a, M_1(a^*), M_2(a)) - Y(a, M_1(a^*), M_2(a^*))\}$ . It is given by,

$$\begin{aligned} & Y(a, M_1(a), M_2(a)) - Y(a, M_1(a), M_2(a^*)) - Y(a, M_1(a^*), M_2(a)) + Y(a, M_1(a^*), M_2(a^*)) \\ &= \sum_{m_1} \sum_{m_2} Y(a, m_1, m_2) \{ I(M_1(a)=m_1) I(M_2(a)=m_2) - I(M_1(a)=m_1) I(M_2(a^*)=m_2) \\ &\quad - I(M_1(a^*)=m_1) I(M_2(a)=m_2) + I(M_1(a^*)=m_1) I(M_2(a^*)=m_2) \} \\ &= \sum_{m_1} \sum_{m_2} Y(a, m_1, m_2) [ I(M_1(a)=m_1) \{ I(M_2(a)=m_2) - I(M_2(a^*)=m_2) \} \\ &\quad - I(M_1(a^*)=m_1) \{ I(M_2(a)=m_2) - I(M_2(a^*)=m_2) \} ] \\ &= \sum_{m_1} \sum_{m_2} Y(a, m_1, m_2) \{ I(M_1(a)=m_1) - I(M_1(a^*)=m_1) \} \{ I(M_2(a)=m_2) - I(M_2(a^*)=m_2) \}. \end{aligned}$$

$$(14)$$

By taking the conditional expectation of (14) given  $C = c$ , we obtain:

$$\begin{aligned}
 & E[\sum_{m_1} \sum_{m_2} Y(a, m_1, m_2) \{I(M_1(a)=m_1) - I(M_1(a^*)=m_1)\} \{I(M_2(a)=m_2) - I(M_2(a^*)=m_2)\} | c] \\
 &= \sum_{m_1} \sum_{m_2} E[Y(a, m_1, m_2) \{I(M_1(a)=m_1) - I(M_1(a^*)=m_1)\} \{I(M_2(a)=m_2) - I(M_2(a^*)=m_2)\} | c] \\
 &= \sum_{m_1} \sum_{m_2} E[Y(a, m_1, m_2) | c] E[I(M_1(a)=m_1) - I(M_1(a^*)=m_1) | c] \\
 &\quad \times E[I(M_2(a)=m_2) - I(M_2(a^*)=m_2) | c] \\
 &= \sum_{m_1} \sum_{m_2} E[Y(a, m_1, m_2) | c] \{Pr(M_1(a)=m_1 | c) - Pr(M_1(a^*)=m_1 | c)\} \\
 &\quad \times \{Pr(M_2(a)=m_2 | c) - Pr(M_2(a^*)=m_2 | c)\},
 \end{aligned}$$

where the second equality follows from B4.

In summary, at the population level, we obtain the following decomposition for the total effect:

$$\begin{aligned}
 E[Y(a) - Y(a^*) | c] &= E[Y(a, M_1(a^*), M_2(a^*)) - Y(a^*, M_1(a^*), M_2(a^*)) | c] \\
 &\quad + E[Y(a, M_1(a), M_2(a^*)) - Y(a, M_1(a^*), M_2(a^*)) | c] \\
 &\quad + E[Y(a, M_1(a^*), M_2(a)) - Y(a, M_1(a^*), M_2(a^*)) | c] \\
 &\quad + \sum_{m_1} \sum_{m_2} E[Y(a, m_1, m_2) | c] \{Pr(M_1(a)=m_1 | c) - Pr(M_1(a^*)=m_1 | c)\} \\
 &\quad \times \{Pr(M_2(a)=m_2 | c) - Pr(M_2(a^*)=m_2 | c)\}, \tag{15}
 \end{aligned}$$

where the first term in (15) is the joint pure natural direct effect, the second term is the indirect effect through  $M_1$  only, and the third term is the indirect effect through  $M_2$  only, and the fourth term is the mediated interactive effect between  $M_1$  and  $M_2$ .

The identification formula of the first three components of (15) are given in (9) of the main text. Using the results in Section 5, the general identification formula for the fourth component is given as follows:

$$\begin{aligned}
 & \sum_{m_1} \sum_{m_2} E[Y(a, m_1, m_2) | c] \{Pr(M_1(a)=m_1 | c) - Pr(M_1(a^*)=m_1 | c)\} \\
 &\quad \times \{Pr(M_2(a)=m_2 | c) - Pr(M_2(a^*)=m_2 | c)\} \\
 &= \sum_{m_1} \sum_{m_2} E[Y | a, m_1, m_2, c] \{p(m_1 | a, c) - p(m_1 | a^*, c)\} \{p(m_2 | a, c) - p(m_2 | a^*, c)\}.
 \end{aligned}$$

### Appendix 3. Derivation of (8)

Under assumptions B1–B4, we have:

$$\begin{aligned}
 E[Y(a, M_1(a^*), M_2(a^{**})) | c] &= \sum_{m_1} \sum_{m_2} E[Y(a, m_1, m_2) | M_1(a^*)=m_1, M_2(a^{**})=m_2, c] Pr(M_1(a^*)=m_1 | c) Pr(M_2(a^{**})=m_2 | c) \\
 &= \sum_{m_1} \sum_{m_2} E[Y(a, m_1, m_2) | c] p(m_1 | a^*, c) p(m_2 | a^{**}, c) \\
 &= \sum_{m_1} \sum_{m_2} E[Y | a, m_1, m_2, c] p(m_1 | a^*, c) p(m_2 | a^{**}, c),
 \end{aligned}$$

where the first equality follows from the law of conditional expectation and B4, the second from B3 and B4, and the third from B1 and B2.

### Appendix 4. Asymptotic distribution of the IPW estimator

We use large sample theory to show that the IPW estimator is asymptotically normal and derive the asymptotic variance using the standard theory of M-estimation.<sup>38</sup> This approach

uses models for  $\Pr(A = a | c) = \pi(a | c; \gamma)$ ,  $\Pr(M_1 = m_1 | a, c) = \xi_1(m_1 | a, c; \beta^{(1)})$ , and  $\Pr(M_2 = m_2 | a, c) = \xi_2(m_2 | a, c; \beta^{(2)})$ . We assume these probabilities are estimated by the maximum likelihood method. We first derive the influence function of the IPW estimator for  $E[Y(a, M_1(a^*), M_2(a^{**}))] = \mu(a, a^*, a^{**}) = \mu$ . Let  $\beta^T = (\beta^{(1)T}, \beta^{(2)T})^T$  and let  $U(\mu, \beta, \gamma) = U(\delta)$  be the estimating function corresponding to the IPW estimator:

$$U(\delta) = \frac{I(A=a)}{\pi(a|c;\gamma)} \times \frac{\xi_1(m_1|a^*, c; \beta^{(1)})}{\xi_1(m_1|a, c; \beta^{(1)})} \times \frac{\xi_2(m_2|a^{**}, c; \beta^{(2)})}{\xi_2(m_2|a, c; \beta^{(2)})} \times (Y - \mu).$$

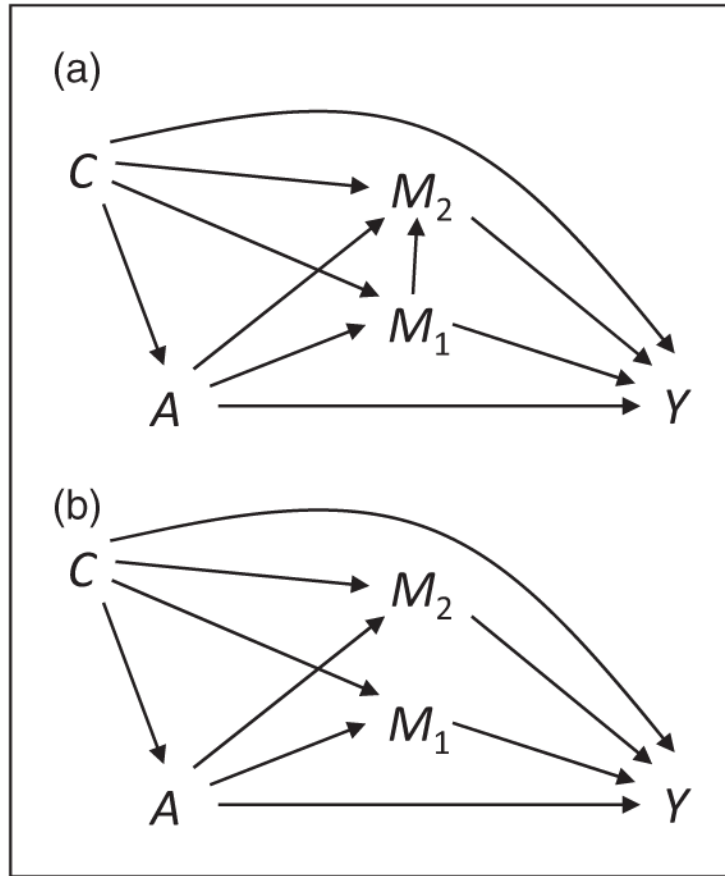
Then, it follows from a Taylor series expansion that

$$\begin{aligned} n^{-1/2} \sum_{i=1}^n U_i(\hat{\delta}) &= n^{-1/2} \sum_{i=1}^n U_i(\delta_0) - n^{1/2}(\hat{\mu}_{IPW} - \mu_0) + H_\beta(\delta_0)n^{1/2}(\hat{\beta} - \beta_0) + H_\gamma(\delta_0)n^{1/2}(\hat{\gamma} - \gamma_0) + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^n U_i(\delta_0) - n^{1/2}(\hat{\mu}_{IPW} - \mu_0) + n^{-1/2}H_\beta(\delta_0)I_\beta^{-1}(\beta_0) \sum_{i=1}^n S_{\beta,i}(\beta_0) + n^{-1/2}H_\gamma(\delta_0)I_\gamma^{-1}(\gamma_0) \sum_{i=1}^n S_{\gamma,i}(\gamma_0) + o_p(1), \end{aligned} \tag{16}$$

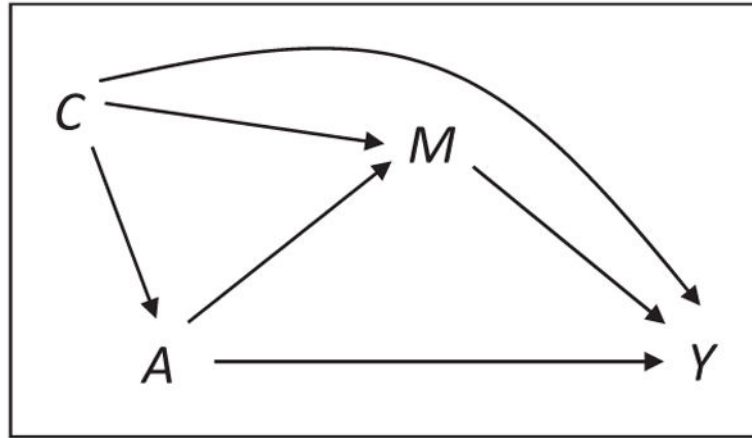
where  $\delta_0 = (\mu_0, \beta_0, \gamma_0)$  are the probability limits of these estimators,  $H_\beta(\delta) = E[U(\delta)/\beta^T]$ ,  $H_\gamma(\delta) = E[U(\delta)/\gamma^T]$ ,  $S_{\beta,i}(\cdot)$  and  $S_{\gamma,i}(\cdot)$  are the corresponding score functions and Fisher information matrices for  $\beta$  and  $\gamma$ . Using  $n^{-1/2} \sum_{i=1}^n U_i(\hat{\delta}) = 0$  and (16), we obtain,

$$n^{1/2}(\hat{\mu}_{IPW} - \mu_0) = n^{-1/2} \sum_{i=1}^n \psi_i + o_p(1), \tag{17}$$

where  $\psi_i = U_i(\delta_0) + H_\beta(\delta_0)I_\beta^{-1}(\beta_0)S_{\beta,i}(\beta_0) + H_\gamma(\delta_0)I_\gamma^{-1}(\gamma_0)S_{\gamma,i}(\gamma_0)$ . From the representation (17), we say that  $n^{1/2}(\hat{\mu}_{IPW} - \mu_0)$  is regular and asymptotically linear with influence function  $\psi_i$ . By replacing the unknown quantities in equation (17) with estimators, we obtain the estimated influence function  $\hat{\psi}_i$ . It now follows from application of the central limit theorem to the representation (17),  $n^{1/2}(\hat{\mu}_{IPW} - \mu_0)$  has an asymptotically normal distribution with mean zero and variance given by  $\text{var}(\psi_i)$ . The asymptotic variance estimator is obtained by calculating the sample variance of  $\hat{\psi}_i$ . For calculating the asymptotic variance estimators of  $\hat{E}[PSE_1(0)]$ ,  $\hat{E}[PSE_2(0)]$ , and  $\hat{E}[MI]$ , we only need to calculate the sample variances of  $\{\hat{\psi}_i(1, 1, 0) - \hat{\psi}_i(1, 0, 0)\}$ ,  $\{\hat{\psi}_i(1, 0, 1) - \hat{\psi}_i(1, 0, 0)\}$ , and  $\{\hat{\psi}_i(1, 1, 1) - \hat{\psi}_i(1, 1, 0) - \hat{\psi}_i(1, 0, 1) + \hat{\psi}_i(1, 0, 0)\}$ , respectively, where  $\hat{\psi}_i(a, a^*, a^{**})$  is the estimated influence function of the IPW estimator for  $\mu(a, a^*, a^{**})$ .

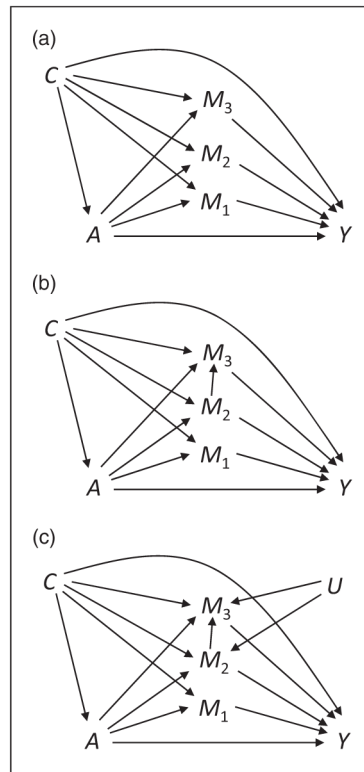


**Figure 1.** A causal diagram with treatment  $A$ , mediators  $M_1$  and  $M_2$ , outcome  $Y$ , and confounding factors  $C$  under (a)  $M_1$  causally affects  $M_2$  and (b)  $M_1$  does not causally affect  $M_2$ .



**Figure 2.** A causal diagram in a single mediator setting with treatment  $A$ , mediator  $M$ , outcome  $Y$ , and confounding factors  $C$ .





**Figure 3.** A causal diagram with treatment  $A$ , mediators  $M_1$ ,  $M_2$ , and  $M_3$ , outcome  $Y$ , and confounding factors  $C$  under (a) three mediators are causally non-ordered, (b)  $M_2$  causally affects  $M_3$ , and (c) there is unmeasured common cause  $U$  of  $M_2$  and  $M_3$ .

**Table 1**

Subjects characteristics at baseline in the CAMBRA trial.

Variables	Intervention ( <i>n</i> = 51)	Control ( <i>n</i> = 50)
Age (years, mean $\pm$ SD)	40.2 $\pm$ 15.1	41.3 $\pm$ 14.9
Female ( <i>n</i> , %)	35, 68.6	31, 62.0
Race ( <i>n</i> , %)		
Asian, non-Hispanic	13, 25.5	9, 18.0
Black, non-Hispanic	7, 15.7	7, 14.0
White, non-Hispanic	8, 13.7	17, 34.0
Hispanic and others	23, 45.1	17, 34.0
Education ( <i>n</i> , %)		
High school	27, 53.0	25, 50.0
Collage	15, 29.4	15, 30.0
Graduate/professional	9, 17.7	10, 20.0
Last dental visit ( <i>n</i> , %)		
Less than 1 year	24, 47.1	19, 38.0
2 to 3 years	15, 29.4	20, 40.0
3 years or more	12, 23.5	11, 22.0
Brushed 2 or more times yesterday ( <i>n</i> , %)	38, 74.5	39, 78.0
Used fluoride toothpaste ( <i>n</i> , %)	37, 72.6	38, 76.0
Fair or poor oral health ( <i>n</i> , %)	19, 37.3	21, 42.0
Drank alcohol in past week ( <i>n</i> , %)	26, 51.0	23, 46.0
Smoked cigarette within 30 days ( <i>n</i> , %)	7, 13.7	6, 12.0

CAMBRA: Caries Management by Risk Assessment; SD: standard deviation.

**Table 2**

Natural direct, indirect, and total effects of the intervention on the DMFS increment in the CAMBRA trial.

		Intervention ( $n = 51$ ), Mean $\pm$ SD	Control ( $n = 50$ ), Mean $\pm$ SD
Baseline	Log <sub>10</sub> MS	4.26 $\pm$ 1.50	4.48 $\pm$ 1.30
	Log <sub>10</sub> LB	3.72 $\pm$ 1.93	3.71 $\pm$ 1.99
	Log <sub>10</sub> fluoride (ppm)	-1.59 $\pm$ 0.22	-1.62 $\pm$ 0.21
12 months	Log <sub>10</sub> MS	3.25 $\pm$ 1.97	4.58 $\pm$ 1.49
	Log <sub>10</sub> LB	2.92 $\pm$ 2.12	3.28 $\pm$ 2.00
	Log <sub>10</sub> fluoride (ppm)	-1.38 $\pm$ 0.44	-1.54 $\pm$ 0.27
24 months	DMFS increment	3.63 $\pm$ 3.54	4.46 $\pm$ 4.16
	Joint natural direct effect (95% CI)	-0.298 (-1.894, 1.805)	
	Joint natural indirect effect (95% CI)	-0.490 (-1.652, 0.172)	
	$M_1$	-0.373 (-1.541, 0.195)	
	$M_2$	-0.022 (-0.366, 0.789)	
	Mediated interaction	-0.095 (-0.807, 0.171)	
	Total effect	-0.788 (-2.108, 0.847)	

CAMBRA: Caries Management by Risk Assessment; CI: confidence interval; DMFS: decayed missing filled permanent surfaces; LB: lactobacilli; MS: mutans streptococci; SD: standard deviation.