Causal Inference with Intermediates: Simple Methods for Principal Strata Effects and Natural Direct Effects

Yasutaka Chiba and Etsuji Suzuki

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/53193

1. Introduction

A central problem in natural science is identifying general laws of cause and effect. Medical science is devoted to revealing causal relationships in humans [1]. The framework for causal inference applied in epidemiology can contribute substantially to clearly specifying and testing causal hypotheses. In some situations, conditioning on an intermediate, which may be between the cause (exposure) and effect (outcome), is of concern for biomedical researchers and public health practitioners [2-4]. In particular, there is a conflict in the perinatal epidemiology literature between the desire to obtain birth-weight-specific associations [5-7] and increasing awareness that conditioning on this variable can give rise to severe biases [8-11]. The difficulty arises because birth weight may be on a pathway from the exposure of interest to the perinatal outcome. For example, if the exposure is maternal smoking and the outcome is infant mortality, maternal smoking may partly affect infant mortality through its effects on fetal growth or on the timing of delivery, thereby potentially through the intermediate, birth weight. In an analysis conditioned on an intermediate, without controlling for the common causes of the intermediate and the outcome, biased results and paradoxical findings can emerge [2-4,12].

It has been reported that maternal smoking appears to have a protective effect against infant mortality among infants with low birth weight [13-15]. This perplexing association is often referred to as the *birth-weight paradox*. This relationship is exemplified by data from cohort-linked birth certificate and infant mortality files for 1997 from the National Center for Health Statistics (NCHS), which are complete files for all US births in 1997 with a 1-year follow-up for infant mortality. Table 1 gives infant mortality statistics stratified by smoking and low-birth weight [16]. In this table, only singletons are included, smoking status is dichotomized (any smoking during pregnancy versus none), and low birth weight is defined as a birth weight of less than 2,500 g. In the whole population, the crude risk difference was 9.9 – 5.9 = 4.0 (95%

confidence interval [CI]: 3.7, 4.4) per 1,000 live births. However, when we stratified the analysis by birth weight (a potential intermediate between maternal smoking and infant mortality), the risk difference was 4.9 – 2.4 = 2.5 (95% CI: 2.3, 2.7) per 1,000 live births in the subpopulation with birth weight ≥ 2,500 g, whereas it was 51.5 – 64.1 = –12.6 (95% CI: –15.0, –10.1) per 1,000 live births in the subpopulation with birth weight < 2,500 g, seemingly illustrating the birth-weight paradox.

		Infant mortality		
		Live birth	Infant death	Deaths per 1,000
Birth weight ≥ 2,500 g[a]	Smoker	353,335	1,729	4.9
	Non-smoker	2,453,633	5,838	2.4
Birth weight < 2,500 g[b]	Smoker	40,383	2,192	51.5
	Non-smoker	137,154	9,387	64.1
Overall[c]	Smoker	393,830	3,950	9.9
	Non-smoker	2,591,452	15,384	5.9
Total		3,749,676	23,693	6.3

[a]Missing information on 40,747 women.

[b]Missing information on 727,384 women.

[c]Missing information on 768,753 women.

Table 1. Infant mortality (number of deaths per 1,000 live births) among women with singleton pregnancies in the 1997 cohort-linked birth certificate infant mortality files from the National Center for Health Statistics, by birth weight and smoking status

The apparent protective effect of maternal smoking among low-birth-weight infants is an artifact of conditioning on an intermediate without adequate control for intermediate-outcome confounding [8]. In the birth-weight paradox, in addition to maternal smoking, birth defects can be a cause of both low birth weight and infant mortality (Figure 1), but birth defects were not considered in the analysis. For mothers who are smokers and have low-birth-weight infants, the low birth weight could either be a consequence of smoking or a birth defect. For mothers who are non-smokers and have low-birth-weight infants, the low birth weight cannot be a consequence of smoking, and some other cause must be operating [8]. Thus, a comparison of smoking and non-smoking mothers without controlling for birth defects will artificially bias the comparison. For this group of low-birth-weight infants, no smoking and low birth weight occurring together is more likely to be associated with the presence of a birth defect. This form of bias is sometimes referred to as collider-stratification bias [17-20] because, on the path $A \rightarrow M \leftarrow U \rightarrow Y$, two arrows collide at node M. The intermediate variable does not need to have an effect on the outcome for such bias to occur (*i.e.,* the dashed arrow from M to Y in Figure 1 is absent); all that is necessary is for the exposure to affect the intermediate and for there to be an unmeasured common cause of the intermediate and the outcome. Unfortunately, intermediate-outcome confounding cannot be eliminated even when the exposure is randomized.

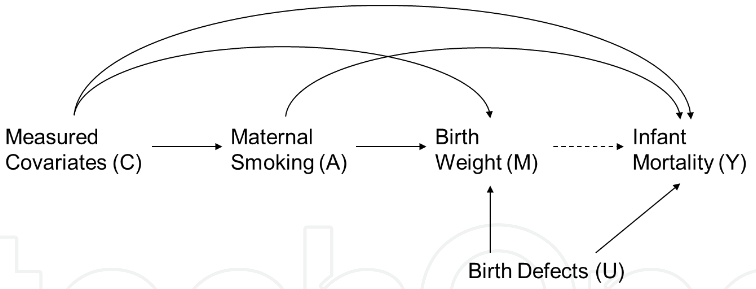

Figure 1. Diagram illustrating the relationships among an exposure (smoking status: *A*), an intermediate (birth weight: *M*), an outcome (infant mortality: *Y*), and both measured (*C*) and unmeasured (*U*) confounders

A considerable volume of literature highlights the hazards of conditioning on an intermediate [9,10,21], but the only solution offered to date is to abandon conditioning on the intermediate altogether. Simply not conditioning on an intermediate will often be the correct way to proceed with an analysis. When the total effect of the exposure on the outcome is of interest, there is no reason to condition on an intermediate. In general, conditioning on an intermediate will be a concern only when other types of effect, such as the direct effect of the exposure on the outcome (not acting through the intermediate), are considered.

In this chapter, we discuss two analytical approaches to help draw inferences when the effect of interest may be obtained by conditioning on an intermediate. The two approaches include (i) the principal stratification approach to assess the *principal strata effect* (PSE), which is a causal effect conditioned on the subpopulation of individuals for whom the intermediate would occur irrespective of exposure status, and (ii) the intervention-based approach to assess the *natural direct effect* (NDE), which is a causal effect capturing what would be realized if the exposure were to occur and its effect on the intermediate were somehow blocked. Each approach has a different interpretation and different methods for the inference. When sensitivity analysis techniques are used for the inference, a range of estimates, rather than a single estimate, will be obtained. As will be discussed later, the two approaches estimate different causal effects, and the resulting estimates would not be expected to be the same. We illustrate each by applying it to the NCHS data in Table 1, causing the birth-weight paradox to evaporate. The approaches are applicable to a variety of similar settings in all areas of epidemiological research.

This chapter is organized as follows. Section 2 introduces the notation used throughout the chapter. Section 3 defines the PSE and presents a simple method for sensitivity analysis. The method is illustrated using the NCHS data. In a similar manner, the NDE is discussed in Section 4. In Section 5, the relationship between these two causal effects is briefly discussed, although this may be somewhat theoretical. Finally, Section 6 offers some concluding remarks.

2. Notation and concepts

We let A denote the exposure of interest (*i.e.*, maternal smoking status), with $A = 1$ for a smoking mother and $A = 0$ for a non-smoking mother. M is the intermediate (*i.e.*, birth weight), with $M = 1$ for a low birth weight (< 2,500 g) and $M = 0$ for a higher birth weight (≥ 2,500 g). Y is the outcome of interest (*i.e.*, infant mortality), with $Y = 1$ for an infant death and $Y = 0$ for a live birth. We let C denote a set of baseline characteristics measured prior to or concurrent with the exposure, and U denote a set of unmeasured confounders between the intermediate (M) and the outcome (Y) (*e.g.*, birth defects). The relationship among the variables is depicted in Figure 1. We assume that no confounder exists between A and M or between A and Y (in Figure 1, a direct arrow from C to A is removed), although this assumption is not practical in the current setting[1]. Nevertheless, even in cases in which confounders exist between these variables, theories presented in this chapter hold conditional on the confounders (C).

Using the above notation, a comparison of the infant mortality risks between smoking and non-smoking mothers with infants of low birth weight can be described as follows:

$$\mathrm{E}[Y \mid A=1, M=1]-\mathrm{E}[Y \mid A=0, M=1],$$

where $\mathrm{E}[Y \mid A = 1, M = 1]$ is estimated by the sample mean of Y (*i.e.*, infant mortality risk) among smoking mothers whose infants had a low birth weight (51.5/1,000) and $\mathrm{E}[Y \mid A = 0, M = 1]$ is estimated by the sample mean of Y among non-smoking mothers whose infants had a low birth weight (64.1/1,000). Although simple, this comparison is not a fair comparison because it compares outcomes for different populations, rather than for the same population with respect to the effect of maternal smoking status on infant mortality. In the second population with $(A, M) = (0, 1)$, even if mothers are smokers, the infants may still have a low birth weight. However, in the first population with $(A, M) = (1, 1)$, if the mothers are non-smokers, some infants may not have a low birth weight. The subpopulation in which infants may have a low birth weight from smoking mothers but not from non-smoking mothers is included in the sample mean when examining smoking mothers whose infants have a low birth weight, but would not be included in the sample mean when examining non-smoking mothers whose infants have a low birth weight. To fairly compare the effect of maternal smoking status in the same population, we need an alternative to the measure described above.

	Potential outcomes		
Response type	Y(1)	Y(0)	Description
1	1	1	Always birth death
2	1	0	Birth death only with maternal smoking
3	0	1	Birth death only with maternal non-smoking
4	0	0	Never birth death

Table 2. Response types for outcome *Y* (*i.e.*, infant mortality) and corresponding potential outcomes

To illustrate the ideas, we use the concept of potential (or counterfactual) outcomes [22,23]. We let $Y(a)$ denote the potential outcome for each individual if, possibly contrary to fact, the exposure were set to level a. When A is binary, the two possible potential outcomes for each individual are $Y(1)$ and $Y(0)$, which respectively correspond to the outcomes that would have happened to the individual with and without maternal smoking. Note that for each individual,

[1] Unfortunately we do not have access to the full NCHS dataset and can only use the published data [16]. Therefore, we cannot implement any analyses in which the confounders are taken into account.

only $Y(1)$ or $Y(0)$ can be observed, depending on the actual exposure status. As a result, individuals can be classified into four different response types, as enumerated in Table 2. Based on these potential outcomes, we can describe the population risk of Y in the presence of A as E[$Y(1)$] and that in the absence of A as E[$Y(0)$]. Thus, a causal risk difference in the whole population is given by

$$\mathrm{E}[Y(1)]-\mathrm{E}[Y(0)].$$

In other words, the above measure quantifies the total effect of A on Y. Under the assumption that no confounder exists between A and Y, one can infer a causal effect of A on Y by simply comparing the observed sample mean of Y among smoking mothers (*i.e.*, $\mathrm{E}[Y \mid A = 1]$) and the observed sample mean of Y among non-smoking mothers (*i.e.*, $\mathrm{E}[Y \mid A = 0]$). Thus, the causal risk difference can be estimated by calculating an associational risk difference as follows:

$$\mathrm{E}[Y \mid A=1]-\mathrm{E}[Y \mid A=0].$$

In the NCHS data, this measure is calculated as 9.9 – 5.9 = 4.0 per 1,000 live births.

Likewise, we let $M(a)$ denote the potential intermediate for each individual if, possibly contrary to fact, the exposure were set to level a. Thus, there are two possible potential intermediates, $M(1)$ and $M(0)$, resulting in four different response types (Table 3). The concept of PSE is closely related to these response types, as discussed in the following section.

	Potential intermediates		
Response type	**M(1)**	**M(0)**	**Description**
1	1	1	Always low birth weight
2	1	0	Low birth weight only with maternal smoking
3	0	1	Low birth weight only with maternal non-smoking
4	0	0	Never low birth weight

Table 3. Response types for intermediate M (*i.e.*, birth weight) and corresponding potential intermediates

Finally, we let $Y(a,m)$ denote the potential outcome for each individual if A were set to a and M were set to m. In this setting, there would be four potential outcomes for each individual, resulting in 4 × 4 = 16 possible response types [24,25]. Using the concept of these potential outcomes, PSE and NDE will be defined in Sections 3 and 4.

3. Principal stratification approach

As an alternative to the crude measure, we introduce the principal stratification approach. We define the PSE in Section 3.1 and present a simple method for the sensitivity analysis under the assumption in Section 3.2. In Section 3.3, the method is illustrated using the NCHS data. In Section 3.4, we present a sensitivity analysis formula by relaxing the assumption used in

Section 3.2, although the form may not be simple. The derivations of the equations and inequalities presented in this section are given in Appendix 1.

3.1. Principal strata effect

One approach to making a fair comparison involves assessing the effect of the exposure on the outcome among the subpopulation for which the intermediate would be present irrespective of the exposure status. For example, we might be interested in the effect among the subpopulation for which infants would have a low birth weight irrespective of maternal smoking status. This subpopulation for which the intermediate will occur irrespective of the exposure is sometimes referred to as a principal stratum [26]. More generally, a principal stratum is a subpopulation defined by the joint potential intermediates ($M(0)$, $M(1)$). If exposure A and intermediate M are binary, there are four possible principal strata:

i. Those for which the intermediate will occur irrespective of exposure status: always low birth weight (response type 1 in Table 3);

ii. Those for which the intermediate will occur with exposure but not without exposure: low birth weight only with maternal smoking (response type 2 in Table 3);

iii. Those for which the intermediate will occur without exposure but not with exposure: low birth weight only with maternal non-smoking or defiers (response type 3 in Table 3); and

iv. Those for which the intermediate will not occur irrespective of exposure status: never low birth weight (response type 4 in Table 3).

If we are interested in whether maternal smoking has a protective effect among low-birth-weight infants, one potentially relevant question within the context of principal stratification is whether maternal smoking has a protective effect among the subpopulation in which infants would have a low birth weight irrespective of maternal smoking status ($M(0) = 1$, $M(1) = 1$). This effect is referred to as a PSE or a principal stratum direct effect, and is formalized as follows [27,28]:

$$\mathrm{PSE} \equiv \mathrm{E}[Y(1) - Y(0) \mid M(0) = 1,\ M(1) = 1].$$

This measure quantifies the total effect of A on Y among the subpopulation defined by the response type of M. The advantage of using the principal stratification approach is that it essentially avoids the problem of conditioning directly on the intermediate. Instead, we condition on the principal stratum, which is essentially an underlying characteristic of the individual. It is like conditioning on a baseline covariate [16]. However, a disadvantage of this approach is that we do not know who is in each principal stratum. For example, we do not know which infants will have a low birth weight irrespective of maternal smoking status. Because we cannot identify the individuals who fall into each principal stratum, we cannot estimate the PSE directly from the observed data. However, the PSE can be estimated from the data by making a number of assumptions [28-30]. Because the PSE cannot be identified when unmeasured confounders exist between the intermediate and the outcome, as shown in Figure

1, some researchers have used sensitivity analysis techniques to assess the magnitude of the PSE [27,31,32].

3.2. Sensitivity analysis method

To derive a simple sensitivity analysis formula, we require the following assumption, which is sometimes referred to as a monotonicity assumption [33]:

Assumption 1. $M(0) \leq M(1)$ for all individuals.

This assumption implies that there are no individuals for whom the intermediate would occur without exposure but not with exposure (*i.e.*, no defiers exist), because $M(1) = 0$ and $M(0) = 1$ cannot hold simultaneously under this assumption. In the context of the smoking-birth weight example, this implies that there is no infant who would have a low birth weight if their mother was a non-smoker, but would not have a low birth weight if their mother was a smoker. When this is the case, the PSE can be expressed as the difference between the crude risk difference and a sensitivity parameter, under Assumption 1 [32]:

$$\text{PSE} = \mathrm{E}[Y \mid A = 1, M = 1] - \mathrm{E}[Y \mid A = 0, M = 1] - \alpha, \tag{1}$$

where the sensitivity parameter α is given by

$\alpha = \mathrm{E}[Y(1) \mid A = 1, M = 1] - \mathrm{E}[Y(1) \mid A = 0, M = 1]$.

The interpretation of this sensitivity parameter is the difference in infant mortality risks under maternal smoking for two subpopulations: the subpopulation with smoking mothers whose infants had a low birth weight, and the subpopulation with non-smoking mothers whose infants had a low birth weight. The parameter is not identified from the observed data.

The sensitivity analysis can be easily conducted. The sensitivity parameter α is set by the investigator according to what is considered plausible. The parameter can be varied over a range of plausible values to examine how conclusions vary according to different parameter values. The confidence interval of the true PSE for a fixed value of α can be obtained simply by subtracting α from the upper and lower confidence limits of the crude risk difference. Therefore, we can graphically display the result of the sensitivity analysis, where the horizontal axis represents the sensitivity parameter and the vertical axis represents the true PSE.

As it may be troublesome to determine the range of α values to examine in some situations, we present a range of values that α can take under some plausible assumptions. These assumptions are straightforward extensions of those developed to assess the total effect of an exposure on the outcome [34-37]. The first assumption, sometimes referred to as the assumption of monotone treatment selection [35-37], is formalized as follows in the current setting:

Assumption 2. $\mathrm{E}[Y(1) \mid A = 1, M = m] \leq \mathrm{E}[Y(1) \mid A = 0, M = m]$ for all m.

This assumption will hold if the subpopulation with $(A, M) = (0, m)$ is less healthy than the subpopulation with $(A, M) = (1, m)$ when the larger value of Y is more harmful. In the context of the smoking-birth weight example, Assumption 2 with $m = 1$ implies that the subpopulation with $(A, M) = (0, 1)$, which consists of low-birth-weight infants with non-smoking mothers, would be a less healthy population than the subpopulation with $(A, M) = (1, 1)$, which consists of low-birth-weight infants with smoking mothers. Assumption 2 with $m = 0$ implies that the subpopulation with $(A, M) = (0, 0)$, which consists of those who do not have a low birth weight with non-smoking mothers, would be a less healthy population than the subpopulation with $(A, M) = (1, 0)$, which consists of those who do not have a low birth weight with smoking mothers.

Assumption 2 with $m = 1$ seems arguably plausible, because in the subpopulation with $(A, M) = (0, 1)$, even if mothers were smokers, infants would have a low birth weight and the mortality risk in this subpopulation would likely be higher than that in the subpopulation with $(A, M) = (1, 1)$. However, Assumption 2 with $m = 0$ may seem less plausible to some investigators. Nevertheless, Assumption 2 with $m = 0$ is still reasonable because under Assumption 1, Assumption 2 with $m = 0$ is equivalent to assuming:

$$\mathrm{E}[Y(1) \mid M(0)=0, M(1)=0] \leq \mathrm{E}[Y(1) \mid M(0)=0, M(1)=1],$$

which implies that the infant mortality risk in the subpopulation consisting of those who would never have a low-birth-weight infant is not more than that in the subpopulation consisting of those who would have a low-birth-weight infant only with a smoking mother. Under the scenario in which the mother is a smoker, this would indeed be the case. Thus, Assumption 2 with $m = 0$ is also arguably reasonable. We note that under Assumption 1, Assumption 2 with $m = 1$ is equivalent to assuming:

$$\mathrm{E}[Y(1) \mid M(0)=0, M(1)=1] \leq \mathrm{E}[Y(1) \mid M(0)=1, M(1)=1],$$

which implies that the infant mortality risk in the subpopulation consisting of those who would have a low-birth-weight infant only with a smoking mother is not more than that in the subpopulation consisting of those who would always have a low-birth-weight infant.

When Assumption 2 holds in addition to Assumption 1, the range of α becomes [32]:

$$\frac{(p_1 - p_0)\{\mathrm{E}[Y \mid A=1, M=0] - \mathrm{E}[Y \mid A=1, M=1]\}}{p_0} \leq \alpha \leq 0,$$

where $p_a = \Pr(M = 1 \mid A = a)$.

The second assumption is sometimes referred to as the assumption of monotone treatment response [34,36,37] and is formalized as follows in the current setting[2]:

Assumption 3. $\mathrm{E}[Y(0) \mid A=a, M=m] \leq \mathrm{E}[Y(1) \mid A=a, M=m]$ for all a and m.

In the context of the smoking-birth weight example, this assumption implies that in the subpopulation with $(A, M) = (a, m)$, the infant mortality risk is higher if the mother was a smoker than if the mother was a non-smoker, which seems reasonable. When, in addition to Assumption 1, Assumption 3 holds, the range of α becomes:

[2] The assumption of monotone treatment response was originally given as an assumption for all individuals; *i.e.*, $Y(0) \leq Y(1)$ for all individuals [34]. Therefore, Assumption 3 is a somewhat weaker assumption than the original one.

$$\left.\begin{Bmatrix}(1-p_0)\mathrm{E}[Y \mid A=0,\ M=0] \\ -(1-p_1)\mathrm{E}[Y \mid A=1,\ M=0] \\ -(p_1-p_0)\mathrm{E}[Y \mid A=1,\ M=1]\end{Bmatrix}\right/ p_0 \le \alpha \le \mathrm{E}[Y \mid A=1,\ M=1]-\mathrm{E}[Y \mid A=0,\ M=1].$$

When it is considered that, in addition to Assumption 1, both Assumptions 2 and 3 hold, we can use a narrower range derived under these two assumptions.

3.3. Illustration

We now apply the principal stratification approach to the NCHS data shown in Table 1. As noted in Section 1, the crude difference in the mortality risk of low-birth-weight infants between smoking and non-smoking mothers was –12.6 (95% CI: –15.0, –10.1) per 1,000 live births, suggesting that maternal smoking has a protective effect against infant mortality for low-birth-weight infants.

To calculate the PSE defined in Section 3.1, we adjust this crude estimate by the sensitivity parameter α. We set the range of α per 1,000 live births to $-22.1 \le \alpha \le -12.6$ because the ranges were calculated as $-22.1 \le \alpha \le 0$ under Assumption 2 and $-84.0 \le \alpha \le -12.6$ under Assumption 3. Figure 2 shows the result of the sensitivity analysis over this range of α, for which the lower and upper limits of the PSE per 1,000 live births were 0.0 (95% CI: –2.4, 2.4) and 9.6 (95% CI: 7.1, 12.0), respectively.

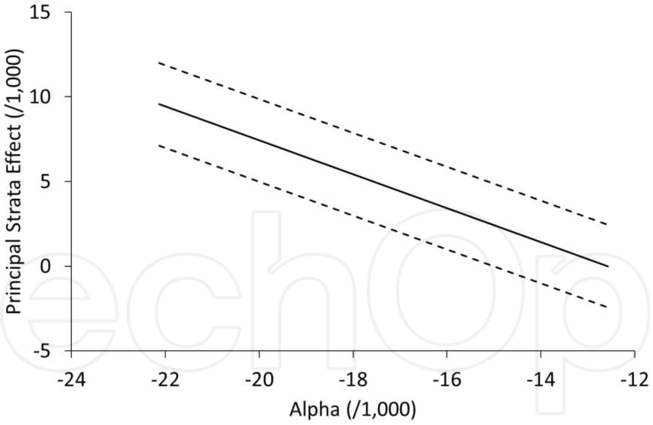

Figure 2. Sensitivity analysis of the principal strata effect (per 1,000 live births); the solid line indicates the principal strata effect and broken lines indicate 95% confidence intervals

The result of this sensitivity analysis for the PSE suggests that maternal smoking has a harmful effect on the subpopulation of infants who would have a low birth weight irrespective of maternal smoking, although the lower limit of the 95% confidence interval for the PSE was still smaller than 0 when α per 1,000 live births was larger than –15.0. Therefore, we can say that the birth-weight paradox was resolved in terms of the PSE.

3.4. Sensitivity analysis without the monotonicity assumption

The monotonicity assumption (Assumption 1) is a strict assumption because the inequality ($M(0) \leq M(1)$) must hold for all individuals. If just one defier exists, this assumption does not hold. Therefore, we introduce a sensitivity analysis for the PSE without the monotonicity assumption. This sensitivity analysis formula requires the following three sensitivity parameters, instead of α:

$$\beta_1 = \mathrm{E}[Y(1) \,|\, M(0)=1,\ M(1)=1] - \mathrm{E}[Y(1) \,|\, M(0)=0,\ M(1)=1],$$

$$\beta_2 = \mathrm{E}[Y(0) \,|\, M(0)=1,\ M(1)=1] - \mathrm{E}[Y(0) \,|\, M(0)=1,\ M(1)=0],$$

$$\beta_3 = \Pr(M(0)=1,\ M(1)=0).$$

In the context of the smoking-birth weight example, the parameter β_1 is the difference in the infant mortality risk under maternal smoking between the subpopulation consisting of those who would always have a low birth weight and the subpopulation consisting of those who would have a low birth weight only with a smoking mother. As discussed in Section 3.2, β_1 will take a positive value. β_2 is the difference in the infant mortality risk under maternal non-smoking between the subpopulation consisting of those who would always have a low birth weight and the subpopulation consisting of those who would have a low birth weight only with a non-smoking mother (defier). β_2 will also take a positive value. β_3 indicates the proportion of defiers. In this context, even if the value of β_3 is not zero, it will be very small.

Using these three sensitivity parameters, the PSE can be expressed as follows [38]:

$$\mathrm{PSE} = \mathrm{E}[Y \mid A=1, M=1] - \mathrm{E}[Y \mid A=0, M=1] + \frac{p_1 - p_0 + \beta_3}{p_1}\beta_1 - \frac{\beta_3}{p_0}\beta_2. \qquad (2)$$

It will be more difficult to determine the values or ranges of these three sensitivity parameters (β_1, β_2, and β_3) compared with those of only one sensitivity parameter (α). Furthermore, it will be difficult to display the result of the sensitivity analysis. For example, in the NCHS data, if we set (β_1, β_2, β_3) = (30.0, 2.0, 0.1) per 1,000 live births, the PSE is 1.7 per 1,000 live births. A larger value of β_1 makes the PSE larger. Conversely, a larger value of β_2 makes the PSE smaller.

4. Intervention-based approach

As another alternative to the crude measure, we introduce the intervention-based approach. In Section 4.1, we define two types of direct effects, the controlled direct effect (CDE) and the NDE, both of which are based on interventions on the intermediate [2,39]. We mainly focus our discussion on the NDE. A simple method for the sensitivity analysis is presented in Section 4.2 and is illustrated in Section 4.3 using the NCHS data. The derivations of equations and inequalities presented in this section are given in Appendix 2.

4.1. Controlled and natural direct effects

The CDE captures the effect of exposure A on outcome Y by intervening to fix intermediate M to m. Using the notation $Y(a,m)$, the CDE is defined as

$$\text{CDE}(m) \equiv \text{E}[Y(1,\ m)] - \text{E}[Y(0,\ m)],$$

Contrary to the PSE, the CDE is a causal effect concerning the whole population. In the context of the smoking-birth weight example, the CDE with m = 1 captures the effect of maternal smoking on infant mortality if all infants were intervened to have a low birth weight; the CDE with m = 0 captures the effect of maternal smoking on infant mortality if all infants were intervened to not have a low birth weight. However, interventions on birth weight seem inconceivable.

The NDE differs from the CDE in that the intermediate M is set to the level $M(0)$, which would have naturally been under the exposure level of $A = 0$. Therefore, to describe the NDE, we need to integrate information about $M(a)$ and $Y(a,m)$. This yields the compound potential outcome $Y(a,M(a^*))$. In this case, individuals can be classified into 4 × 16 = 64 combined response types, each of which has a corresponding pattern of compound potential outcomes [24]. Using the compound potential outcome, the NDE is defined as[3]

$$\text{NDE} \equiv \text{E}[Y(1,\ M(0))] - \text{E}[Y(0,\ M(0))],$$

which compares the effect of an exposure on the outcome if the intermediate were set to what it would have been when exposure A was set to 0. In the context of the smoking-birth weight example, the NDE compares what would have happened to an infant if the mother had been a smoker versus a non-smoker and if the infant had the birth weight status that would have occurred due to maternal non-smoking. Corresponding to the NDE is a natural indirect effect (NIE). The NIE is defined as

$$\text{NIE} \equiv \text{E}[Y(1,\ M(1))] - \text{E}[Y(1,\ M(0))],$$

which compares the effect of the intermediate at levels $M(1)$ and $M(0)$ on the outcome when exposure A is set to 1. The NIE can be interpreted as a causal effect when intervention is made to block a direct link between the exposure and the outcome (an arrow between A and Y in Figure 1), rather than to block a variable itself [39]. Therefore, the NIE is an indirect component of the total effect, acting through the intermediate. Note that the total effect decomposes into the NDE and NIE, *i.e.*, $\text{E}[Y(1)] - \text{E}[Y(0)] = \text{NDE} + \text{NIE}$. This decomposition holds at not only the population level but also the individual level; when we define the individual natural direct and indirect effects by $\text{NDE}(\omega) \equiv Y(1,M(0)) - Y(0,M(0))$ and $\text{NIE}(\omega) \equiv Y(1,M(1)) - Y(1,M(0))$, respectively, for each individual ω,

$$\begin{aligned} Y(1) - Y(0) &= Y(1,\ M(1)) - Y(0,\ M(0)) \\ &= \{Y(1,\ M(1)) - Y(1,\ M(0))\} + \{Y(1,\ M(0)) - Y(0,\ M(0))\} \\ &= \text{NIE}(\omega) + \text{NDE}(\omega). \end{aligned}$$

[3] We can also define the NDE under the exposure level of A = 1 as $\text{NDE} \equiv \text{E}[Y(1,M(1))] - \text{E}[Y(0,M(1))]$. The NIE corresponding to this NDE is equal to $\text{E}[Y(0,M(1))] - \text{E}[Y(0,M(0))]$.

This decomposition may help in understanding the meaning of the NDE, *i.e.*, the NDE is a direct component of the total effect and does not act through the intermediate.

If there is no interaction between the effects of exposure A and intermediate M on outcome Y in the sense that $\mathrm{E}[Y(1,m)] - \mathrm{E}[Y(0,m)]$ does not vary with m, the CDE and NDE coincide. The difference between a total effect and a CDE cannot generally be interpreted as an indirect effect and thus cannot be used to assess mediation. This is because when there is an interaction between the effects of exposure A and intermediate M on outcome Y, the CDEs may differ from the total effect even when A is not a cause of M. When there is an interaction between A and M, the CDEs will differ with different values of m, and thus one of the CDEs will differ from a total effect. Therefore, in general, CDEs cannot be used for decomposition of a total effect into direct and indirect effects. We note that the CDE will often be of greater interest in policy [39], and the NDE will often be of greater interest in the evaluation of etiology [39-41].

Similar to the CDE, the NDE is also a causal effect concerning the whole population, and it can be estimated from the data under certain assumptions [42,43]. However, neither the CDE nor the NDE can be identified when an unmeasured confounder exists between the intermediate and the outcome, as in Figure 1. Therefore, sensitivity analysis techniques have been discussed to assess their magnitudes [4,44-48]. In the next subsection, a simple sensitivity analysis method for the NDE is presented. Methods for the CDE are found elsewhere [4,44,47].

4.2. Sensitivity analysis method for the natural direct effect

Although the monotonicity assumption (Assumption 1) was necessary to derive a simple sensitivity analysis formula for the PSE, this assumption is not required to derive one for the NDE.

For each possible value of intermediate M (= 0, 1), we consider the following sensitivity parameter:

$$\gamma_m = \mathrm{E}[Y(1,\ m) \mid A=1,\ M=m] - \mathrm{E}[Y(1,\ m) \mid A=0,\ M=m].$$

The sensitivity parameter $\gamma_1 = \mathrm{E}[Y(1,1) \mid A = 1, M = 1] - \mathrm{E}[Y(1,1) \mid A = 0, M = 1]$ is a contrast of infant mortality risks for two subpopulations. In a manner similar to the sensitivity parameter α for the PSE, the first subpopulation with $(A, M) = (1, 1)$ consists of smoking mothers whose infants had a low birth weight, and the second subpopulation with $(A, M) = (0, 1)$ comprises non-smoking mothers whose infants had a low birth weight. We then consider whether the infants in these two subpopulations would have lived or died if we had intervened to fix mothers to be smokers and infants to have a low birth weight; *i.e.*, we consider $Y(1,1)$. The contrast between the infant mortality risks in these two subpopulations under this particular intervention is our sensitivity parameter, γ_1. This parameter differs from α for the PSE in that it considers two interventions, which fix A to 1 and M to 1, whereas α considers one intervention, which fixes A to 1. As described above, it is not realistic to consider an intervention to fix the birth weight of an infant. This is a disadvantage of using the sensitivity parameter γ_1 in the smoking-birth weight context. Analogously, the other sensitivity parameter, $\gamma_0 = \mathrm{E}[Y(1,0) \mid A = 1, M = 0] - \mathrm{E}[Y(1,0) \mid A = 0, M = 0]$, can be interpreted. The parameters are not identified from the observed data.

We now consider a weighted mean of the sensitivity parameters γ_1 and γ_0, where the respective weights are the probabilities of infants with low birth weights and not from non-smoking mothers, *i.e.*, $\Pr(M = 1 \mid A = 0)$ and $\Pr(M = 0 \mid A = 0)$:

$$\Gamma = \sum_{m=0}^{1} \gamma_m \Pr(M = m \mid A = 0). \tag{3}$$

After calculating the crude risk differences $E[Y \mid A = 1, M = m] - E[Y \mid A = 0, M = m]$ and probabilities $\Pr(M = m \mid A = 0)$ for $m = 0, 1$, the NDE can be expressed as the difference between the weighted means of the two crude risk differences and Γ, *i.e.*,

$$\mathrm{NDE} = \sum_{m=0}^{1} \{E[Y \mid A = 1, M = m] - E[Y \mid A = 0, M = m]\} \Pr(M = m \mid A = 0) - \Gamma. \tag{4}$$

The variance of the first term in equation (4) is calculated by the delta method, $\mathrm{var}(\hat{s}\hat{t}) = \mathrm{var}(\hat{s})\mathrm{var}(\hat{t}) + s^2\mathrm{var}(\hat{t}) + t^2\mathrm{var}(\hat{s})$, where s and t are replaced by the estimates $\hat{s}$ and $\hat{t}$. In a manner similar to the sensitivity analysis for the PSE, the sensitivity analysis for the NDE can be conducted easily. The sensitivity parameters γ_0 and γ_1 are set by the investigator according to what is considered plausible. The parameters can be varied over a range of plausible values to examine how the conclusions change according to the different parameter values. However, to obtain the confidence interval of the true NDE for the fixed values of γ_m, we must calculate not only the variance of the first term in equation (4) but also the variance of Γ, because Γ depends on the probabilities $\Pr(M = m \mid A = 0)$, which must be estimated from the observed data. However, if γ_m were constant across the strata of m, Γ would no longer depend on $\Pr(M = m \mid A = 0)$, and thus we could simply subtract Γ from both limits of the confidence interval for the first term in equation (4) to obtain the confidence interval for the true NDE. Similarly, for a data set large enough that the estimates of $\Pr(M = m \mid A = 0)$ were very precise, the approximate confidence interval for the true NDE could be obtained by subtracting Γ from both limits of the confidence interval for the first term in equation (4).

We can determine the upper limits of γ_m by using the following assumptions, which are similar to Assumptions 2 and 3 for the PSE:

Assumption 2*. $E[Y(1, m) \mid A = 1, M = m] \le E[Y(1, m) \mid A = 0, M = m]$ for all m.

Assumption 3*. $E[Y(0, m) \mid A = a, M = m] \le E[Y(1, m) \mid A = a, M = m]$ for all a and m.

The upper limit of γ_1 is $\gamma_1 \le 0$ under Assumption 2* with $m = 1$ or $\gamma_1 \le E[Y \mid A = 1, M = 1] - E[Y \mid A = 0, M = 1]$ under Assumption 3* with $a = 0$ and $m = 1$. These upper limits are equal to those of $\alpha = E[Y(1) \mid A = 1, M = 1] - E[Y(1) \mid A = 0, M = 1]$ for the sensitivity analysis of the PSE in Section 3.2. Unfortunately, the lower limit of γ_1 cannot be derived under these assumptions, even when Assumption 1 is added, because we are considering interventions on not only exposure A but also intermediate M (see Appendix 2). Furthermore, it is somewhat difficult to interpret these assumptions because of intervention on the intermediate M. The upper limit of γ_0 is $\gamma_0 \le 0$ under Assumption 2* with $m = 0$ or $\gamma_0 \le \mathrm{E}[Y \mid A = 1, M = 0] - \mathrm{E}[Y \mid A = 0, M = 0]$ under Assumption 3* with $a = 0$ and $m = 0$. From these upper limits of γ_m, the upper limit of Γ is calculated using equation (3).

Assumptions 2* and 3* (2 and 3) relate to monotone treatment selection and monotone treatment response regarding the exposure, respectively. We can also make these types of assumptions about the intermediate [49]:

Assumption 4. $\mathrm{E}[Y(a, m) \mid A = a, M = 0] \le \mathrm{E}[Y(a, m) \mid A = a, M = 1]$ for all a and m.

Assumption 5. $\mathrm{E}[Y(a, 0) \mid A = a^*, M = m] \le \mathrm{E}[Y(a, 1) \mid A = a^*, M = m]$ for all a, a^*, and m.

In the context of the smoking-birth weight example, Assumption 4 implies that infants with a low birth weight, representing the subpopulation with $(A, M) = (a, 1)$, would be a less healthy than infants who did not have a low birth weight, representing the subpopulation with $(A, M) = (a, 0)$, where maternal smoking status is common among these two subpopulations. Assumption 5 implies that in the subpopulation with $(A, M) = (a^*, m)$, the infant mortality risk would be higher if infants had a low birth weight than if infants did not have a low birth weight. As this would indeed be the case, Assumptions 4 and 5 seem arguably reasonable. Under these assumptions, although ranges of γ_m cannot be derived, the range of Γ can be derived as follows:

$$-(1-p_0)\{\mathrm{E}[Y \mid A=1, M=1] - \mathrm{E}[Y \mid A=1, M=0]\} \le \Gamma \le p_0\{\mathrm{E}[Y \mid A=1, M=1] - \mathrm{E}[Y \mid A=1, M=0]\}.$$

While Assumptions 2* and 3* can lead to only the upper limit, Assumptions 4 and 5 can lead to both limits. Furthermore, when Assumption 1 is added, under Assumptions 1 and 5, the lower limit of Γ is improved to:

$$\Gamma \ge -(p_1 - p_0)\{\mathrm{E}[Y \mid A=1, M=1] - \mathrm{E}[Y \mid A=1, M=0]\}.$$

However, neither the lower nor the upper limit can be derived under only one of the Assumptions 4 and 5.

4.3. Illustration

We now apply this intervention-based approach to the NCHS data shown in Table 1. To calculate the NDE defined by equation (4), we determine a range of values for Γ. Under Assumptions 2* and 3*, the respective upper limits of γ_0 and γ_1 per 1,000 live births were calculated as $\gamma_0 \le 0$ and $\gamma_1 \le -12.6$. By substituting these upper limits into equation (3), we obtained $\Gamma \le -0.7$ per 1,000 live births. Under Assumptions 4 and 5, the range of Γ per 1,000 live births was calculated as $-44.0 \le \Gamma \le 2.6$. Under Assumptions 1 and 5, the lower limit was improved to $\Gamma \ge -2.4$ per 1,000 live births. Therefore, we set the range of Γ per 1,000 live births as $-2.4 \le \Gamma \le -0.7$. Because the sample size was large, we obtain the approximate confidence interval for the true NDE by subtracting Γ from both limits of the confidence interval for the first term in equation (4). The result of the sensitivity analysis over this range of Γ is shown in Figure 3. With this range of Γ, the lower and upper limits of the NDE per 1,000 live births were 2.4 (95% CI: 2.0, 2.7) and 4.0 (95% CI: 3.7, 4.4), respectively.

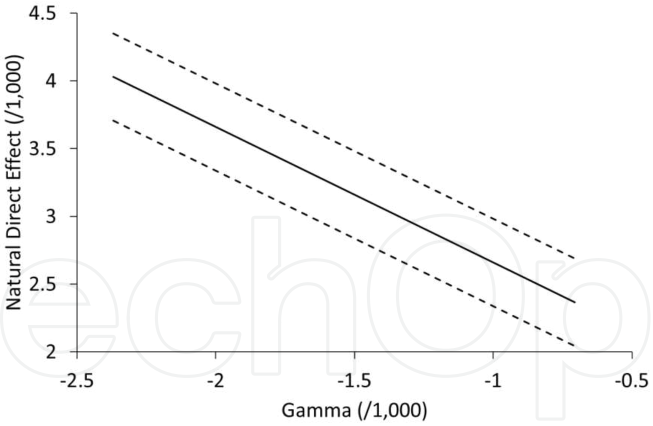

Figure 3. Sensitivity analysis of the natural direct effect (per 1,000 live births); the solid line indicates the natural direct effect and broken lines indicate 95% confidence intervals

The result of this sensitivity analysis for the NDE suggests that maternal smoking has a directly harmful effect on infant mortality. Thus, the birth-weight paradox is also resolved in terms of the NDE.

5. Relationship between the principal strata effect and the natural direct effect

We briefly discuss the relationship between the PSE and NDE. Again, we note that the individual NDE is defined as $\mathrm{NDE}(\omega) \equiv Y(1,M(0)) - Y(0,M(0))$. It can then be shown that, when there is no natural direct effect for any individual, there is no principal strata effect [50], *i.e.*,

Theorem 1. If $\mathrm{NDE}(\omega) = 0$ for all ω, then PSE = 0.

To prove this theorem, we consider a probability of $Y(1) - Y(0) = 0$ conditional on $\{M(0) = 1, M(1) = 1\}$, *i.e.*, $\Pr(Y(1) - Y(0) = 0 \mid M(0) = 1, M(1) = 1)$. This indicates a probability that the PSE is equal to 0. We prove Theorem 1 by showing that this probability is equal to 1 under the assumption that $\mathrm{NDE}(\omega) = 0$ for all ω. The proof is as follows:

$$\begin{aligned}
&\Pr(Y(1)-Y(0)=0 \mid M(0)=1,\ M(1)=1)\\
&=\Pr(Y(1,\ M(1))-Y(1,\ M(0))+\mathrm{NDE}(\omega)=0 \mid M(0)=1,\ M(1)=1)\\
&=\Pr(Y(1,\ M(1))-Y(1,\ M(0))=0 \mid M(0)=1,\ M(1)=1)\\
&=\Pr(Y(1,\ 1)-Y(1,\ 1)=0 \mid M(0)=1,\ M(1)=1)\\
&=\Pr(0=0 \mid M(0)=1,\ M(1)=1)\\
&=1,
\end{aligned}$$

where the first equation is from the decomposition of the total effect to the NDE and NIE, the second is by $\mathrm{NDE}(\omega) = 0$, and the third is because $Y(1,M(a))$ is equal to $Y(1,1)$ conditional on $M(a) = 1$. This completes the proof.

The converse of this theorem does not hold: the absence of a PSE does not imply the absence of a NDE. Nevertheless, from the contraposition of this theorem, when there is a PSE, there must be some individuals for whom there is a NDE. The results of the sensitivity analyses in Sections 3.3 and 4.3 showed that the true PSE was not smaller than 0 and that the true NDE was larger than 2.3. The results do not contradict the contraposition of Theorem 1.

6. Conclusion

In this chapter, we described two approaches related to calculating the effect of an exposure on an outcome that is conditional on potential intermediates or one that does not act through the intermediate. Here, we made an impractical assumption in the observational studies that no confounder exists between the exposure and the outcome or between the exposure and the intermediate. Nevertheless, the methodologies described here also hold conditional on confounders if no unmeasured confounder exists between these variables. We considered a risk difference as the effect measure, but the methodologies can be extended to other effect measures.

Each approach has a unique interpretation and its own strengths and weaknesses. In the principal stratification approach, one conditions on the subpopulation for which the intermediate would occur irrespective of exposure. An advantage of this approach is that the subpopulation is a particularly high-risk group in which the intermediate will necessarily occur. A disadvantage is that we do not know who is in the subpopulation such that the intermediate will occur irrespective of the exposure. In the intervention-based approach, the NDE appears to capture the effect of an exposure on the outcome if the intermediate was set to what it would be when the exposure is set to 0. An advantage of this approach is that it can be used to decompose the total effect into direct and indirect components. A disadvantage is that it is difficult to understand the meaning of the NDE from the form. In addition, it is difficult to interpret the sensitivity parameter for the sensitivity analysis, although this may be avoided by applying a parametric model [46].

In many studies, the total effect of the exposure on the outcome in the whole population may be of central interest, and then none of the approaches described here are required. The approaches described here are of relevance only when the investigators are interested in the direct effect of the exposure not acting through the intermediate or the effect of the exposure on the outcome for certain groups at high risk for the intermediate. In some birth weight settings, the exposure or intervention under study may occur after birth in some cases [6]. In these cases, birth weight becomes a pre-exposure baseline variable, and the approaches described here are not needed. These settings should be distinguished from those similar to the birth-weight paradox. When the approaches described here are of relevance, both the PSE and NDE may be in a consistent direction in some situations, as seen in Sections 3.3 and 4.3. However, it is important to note that the two approaches need not give effect estimates in the same direction. Having effect estimates in different directions with the two approaches is not necessarily an indication that one of the estimates is in the wrong direction. The two approaches estimate two different effects (effects for two different populations), and these may in fact be in different directions. Before these approaches are applied, it is important to be clear about the scientific or policy question.

The approaches described in this chapter are applicable to a number of similar settings in all areas of epidemiological research. As the existing literature has made clear, conditioning on an intermediate can be problematic and can give rise to severe biases. In many contexts, conditioning on an intermediate is not necessary and is best avoided. Nevertheless, there are cases in which conditioning on an intermediate is of scientific or policy interest. We have shown that alternative approaches can be used to draw inferences in such settings. Although these methodological tools are imperfect and need to be interpreted carefully, they can be useful in examining conditional and direct effects.

Appendix

Appendices 1 and 2 outline the derivations of the equations and inequalities presented in Sections 3 and 4, respectively. As noted in Section 2, we assume that no confounder exists between A and M or between A and Y. This assumption leads to the independency assumption that $M(a)$, $Y(a)$, and $Y(a,m)$ are independent of A. In addition, we require two assumptions. The first is the no-interference assumption that one individual's outcome does not depend on the exposure status of other individuals. The second is the consistency assumption that when $A = a$, the potential outcomes $Y(a)$ and $M(a)$ are equal to the observed outcomes Y and M, respectively. Likewise, we assume that when $A = a$ and $M = m$, the potential outcome $Y(a,m)$ is equal to Y. For simplicity, we use the notations $E_{ij}(a) = \mathrm{E}[Y(a) \mid M(0) = i, M(1) = j]$ and $\pi_{ij} \equiv \Pr(M(0) = i, M(1) = j)$.

Appendix 1: Derivations of equations and inequalities in Section 3

Derivation of equation (1)

Using α, PSE can be expressed as follows:

$$\begin{aligned}\mathrm{PSE} &\equiv E_{11}(1) - E_{11}(0)\\ &= \mathrm{E}[Y(1) \mid M(0) = 1] - \mathrm{E}[Y(0) \mid M(0) = 1]\\ &= \mathrm{E}[Y(1) \mid A = 0, M = 1] - \mathrm{E}[Y(0) \mid A = 0, M = 1]\\ &= \{\mathrm{E}[Y(1) \mid A = 1, M = 1] - \alpha\} - \mathrm{E}[Y(0) \mid A = 0, M = 1]\\ &= \mathrm{E}[Y \mid A = 1, M = 1] - \mathrm{E}[Y \mid A = 0, M = 1] - \alpha,\end{aligned}$$

where the second equation is by Assumption 1, the third is by the independency and consistency assumptions, the fourth is by using $\alpha = \mathrm{E}[Y(1) \mid A = 1, M = 1] - \mathrm{E}[Y(1) \mid A = 0, M = 1]$, and the last is again by the consistency assumption.

Proof that Assumption 2 is equivalent to assuming that $E_{01}(1) \leq E_{11}(1)$ and $E_{00}(1) \leq E_{01}(1)$

The relationship between π_{ij} and $p_a = \Pr(M = 1 \mid A = a)$ is

$$\pi_{11} + \pi_{01} = p_1,\ \pi_{10} + \pi_{0_0} = 1 - p_1,\ \pi_{11} + \pi_{1_0} = p_0,\ \ and \pi_{0_1} + \pi_{00} = 1 - p_0 \quad \text{(A1)}$$

because Pr($M(a) = m$) = Pr($M(a) = m \mid A = a$) = Pr($M = m \mid A = a$) by the independency and consistency assumptions. Furthermore, because E[$Y(a) \mid A = 1, M = 1$] = E[$Y(a) \mid M(1) = 1$] by the independency assumption, this conditional expectation can be expressed as

$$\mathrm{E}[Y(a)\,\Big|\,A=1,\ M=1]=\frac{\pi_{11}E_{11}(a)+\pi_{01}E_{01}(a)}{\pi_{11}+\pi_{01}}. \tag{A2}$$

Similarly,

$$\mathrm{E}[Y(a)\,\Big|\,A=0,\ M=1]=\frac{\pi_{11}E_{11}(a)+\pi_{10}E_{10}(a)}{\pi_{11}+\pi_{10}}, \tag{A3}$$

$$\mathrm{E}[Y(a)\,\Big|\,A=1,\ M=0]=\frac{\pi_{10}E_{10}(a)+\pi_{00}E_{00}(a)}{\pi_{10}+\pi_{00}}, \tag{A4}$$

$$\mathrm{E}[Y(a)\,\Big|\,A=0,\ M=0]=\frac{\pi_{01}E_{01}(a)+\pi_{00}E_{00}(a)}{\pi_{01}+\pi_{00}}. \tag{A5}$$

As $\pi_{10} = 0$ (no defier exists) under Assumption 1, (A1) reduces to

$$\pi_{11}=p_0,\ \pi_{00}=\ 1-p_1,\ \text{ and } \pi_{01}=p_1-p_0,$$

where it is assumed that $\pi_{01} > 0$ (*i.e.*, $p_1 > p_0$). Using α = E[$Y(1) \mid A = 1, M = 1$] – E[$Y(1) \mid A = 0, M = 1$] and α' = E[$Y(1) \mid A = 1, M = 0$] – E[$Y(1) \mid A = 0, M = 0$], (A2)–(A5) with $a = 1$ can be expressed respectively as

$$\mathrm{E}[Y\,\Big|\,A=1,\ M=1]=\frac{p_0E_{11}(1)+(p_1-p_0)E_{01}(1)}{p_1}, \tag{A6}$$

$$\mathrm{E}[Y\,|\,A=1,\ M=1]-\alpha=E_{11}(1), \tag{A7}$$

$$\mathrm{E}[Y\,|\,A=1,\ M=0]=E_{00}(1), \tag{A8}$$

$$\mathrm{E}[Y\,\Big|\,A=1,\ M=0]-\alpha'=\frac{(p_1-p_0)E_{01}(1)+(1-p_1)E_{00}(1)}{1-p_0}. \tag{A9}$$

The differences between (A6) and (A7) and between (A8) and (A9) lead to, respectively

$$\alpha=\frac{p_1-p_0}{p_1}\Big\{E_{01}(1)-E_{11}(1)\Big\},$$

$$\alpha'=\frac{p_1-p_0}{1-p_0}\Big\{E_{00}(1)-E_{01}(1)\Big\}.$$

Assumption 2 with $m = 1$ is equivalent to $\alpha \le 0$, and that with $m = 0$ is equivalent to $\alpha' \le 0$. Thus, Assumption 2 is equal to assuming that $E_{01}(1) \le E_{11}(1)$ with $m = 1$ and $E_{00}(1) \le E_{01}(1)$ with $m = 0$, because $p_1 > p_0$ by assumption.

Derivations of ranges of α under Assumptions 2 and 3

Substituting (A7) into (A6) gives

$$E_{01}(1)=\mathrm{E}[Y\,\Big|\,A=1,\ M=1]+\frac{p_0}{p_1-p_0}\alpha, \tag{A10}$$

and substituting (A8) into (A9) leads to

$$E_{01}(1)=\mathrm{E}[Y\,\Big|\,A=1,\ M=0]-\frac{1-p_0}{p_1-p_0}\alpha'. \tag{A11}$$

Given that these two equations are equal, some algebra yields

$$p_0\alpha + (1-p_0)\alpha' = -(p_1 - p_0)\{\mathrm{E}[Y\,|\,A=1,\ M=1] - \mathrm{E}[Y\,|\,A=1,\ M=0]\}. \tag{A12}$$

Then, a range of α under Assumption 2 is derived by $\alpha \le 0$ and by substituting $\alpha' \le 0$ into (A12).

Under Assumption 3,

$$\begin{aligned}\alpha &= \mathrm{E}[Y(1)\,|\,A=1,\ M=1] - \mathrm{E}[Y(1)\,|\,A=0,\ M=1]\\ &\le \mathrm{E}[Y(1)\,|\,A=1,\ M=1] - \mathrm{E}[Y(0)\,|\,A=0,\ M=1]\\ &= \mathrm{E}[Y\,|\,A=1,\ M=1] - \mathrm{E}[Y\,|\,A=0,\ M=1],\end{aligned}$$

and similarly $\alpha' \le \mathrm{E}[Y\,|\,A=1, M=0] - \mathrm{E}[Y\,|\,A=0, M=0]$. Thus, a range of α under Assumption 3 is derived by $\alpha \le \mathrm{E}[Y\,|\,A=1, M=1] - \mathrm{E}[Y\,|\,A=0, M=1]$ and by substituting $\alpha' \le \mathrm{E}[Y\,|\,A=1, M=0] - \mathrm{E}[Y\,|\,A=0, M=0]$ into (A12).

Derivation of equation (2)

Using $\beta_1 = E_{11}(1) - E_{01}(1)$, $\beta_2 = E_{11}(0) - E_{10}(0)$ and $\beta_3 = \pi_{10}$, (A2) with $a = 1$ can be expressed as

$$\begin{aligned}\mathrm{E}[Y\,|\,A=1,\ M=1] &= \frac{(\pi_{11}+\pi_{01})E_{11}(1) - \pi_{01}\{E_{11}(1)-E_{01}(1)\}}{\pi_{11}+\pi_{01}}\\ &= E_{11}(1) - \frac{p_1-(p_0-\pi_{10})}{p_1}\left\{E_{11}(1)-E_{01}(1)\right\}\\ &= E_{11}(1) - \frac{p_1-p_0+\beta_3}{p_1}\beta_1,\end{aligned}$$

and (A3) with $a = 0$ can be expressed similarly as

$$\begin{aligned}\mathrm{E}[Y\,|\,A=0,\ M=1] &= \frac{(\pi_{11}+\pi_{10})E_{11}(0) - \pi_{10}\{E_{11}(0)-E_{10}(0)\}}{\pi_{11}+\pi_{01}}\\ &= E_{11}(0) - \frac{\pi_{10}}{p_0}\left\{E_{11}(0)-E_{10}(0)\right\}\\ &= E_{11}(0) - \frac{\beta_3}{p_0}\beta_2.\end{aligned}$$

The difference between these two equations leads to equation (2).

Appendix 2: Derivations of equations and inequalities in Section 4

Derivation of equation (4)

Using γ_m and Γ, $\mathrm{E}[Y(1,M(0))]$ can be expressed as follows:

$$\begin{aligned}\mathrm{E}[Y(1,\ M(0))] &= \mathrm{E}[Y(1,\ M(0))\,|\,A=0]\\ &= \sum_m \mathrm{E}[Y(1,\ M(0))\,|\,A=0,\ M(0)=m]\Pr(M(0)=m\,|\,A=0)\\ &= \sum_m \mathrm{E}[Y(1,\ m)\,|\,A=0,\ M=m]\Pr(M=m\,|\,A=0)\\ &= \sum_m \{\mathrm{E}[Y(1,\ m)\,|\,A=1,\ M=m]-\gamma_m\}\Pr(M=m\,|\,A=0)\\ &= \sum_m \mathrm{E}[Y\,|\,A=1,\ M=m]\Pr(M=m\,|\,A=0)-\Gamma,\end{aligned}$$

where the first equation is by the independency assumption, the third is by the consistency assumption, and the fourth is by $\gamma_m = \mathrm{E}[Y(1,m)\,|\,A=1, M=m] - \mathrm{E}[Y(1,m)\,|\,A=0, M=m]$. The simpler calculation expresses $\mathrm{E}[Y(0,M(0))]$ as

$$\begin{aligned}\mathrm{E}[Y(0,\ M(0))] &= \mathrm{E}[Y(0)]\\ &= \mathrm{E}[Y \mid A=0]\\ &= \sum_m \mathrm{E}[Y \mid A=0,\ M=m]\Pr(M=m \mid A=0).\end{aligned}$$

The difference between these two equations leads to equation (4).

The reason that the lower limit of γ_{m} cannot be derived under Assumptions 2* and 3*

A range of α can be derived because (A10) and (A11) are equal. In the setting in which γ_m is used instead of α and α', under Assumption 1, the following equations are derived, instead of (A10) and (A11):

$$E_{01}(1,\ 1) = \mathrm{E}[Y \mid A=1,\ M=1] + \frac{p_0}{p_1 - p_0}\gamma_1, \tag{A13}$$

$$E_{01}(1,\ 0) = \mathrm{E}[Y \mid A=1,\ M=0] - \frac{1-p_0}{p_1 - p_0}\gamma_0, \tag{A14}$$

where $E_{ij}(a,m)$ = E[$Y(a,m)$ | $M(0) = i$, $M(1) = j$]. Because these two equations are not equal, the lower limit of γ_0 (γ_1) cannot be derived using the upper limit of γ_1 (γ_0) under Assumptions 2* and 3*, even under Assumption 1.

Derivations of ranges of Γ under Assumptions 4 and 5 and Assumptions 1 and 5

By the consistency assumption, E[$Y(1,m)$ | A = 1, M = m] = E[Y | A = 1, M = m]. Using Assumptions 4 and 5, the following inequality can be derived:

$$\begin{aligned}\sum_m \mathrm{E}[Y(1,\ m) \mid A=0,\ M=m]\Pr(M=m \mid A=0) &\le \sum_m \mathrm{E}[Y(1,\ 1) \mid A=0,\ M=m]\Pr(M=m \mid A=0)\\ &= \mathrm{E}[Y(1,\ 1) \mid A=0]\\ &= \mathrm{E}[Y(1,\ 1) \mid A=1]\\ &= \sum_m \mathrm{E}[Y(1,\ 1) \mid A=1,\ M=m]\Pr(M=m \mid A=1)\\ &\le \sum_m \mathrm{E}[Y(1,\ 1) \mid A=1,\ M=1]\Pr(M=m \mid A=1)\\ &= \mathrm{E}[Y \mid A=1,\ M=1],\end{aligned}$$

where the first inequality is by Assumption 5 with a = 1 and a^* = 0, the third equation is by the independency assumption, and the fifth inequality is by Assumption 4 with a = 1 and m = 1. Substituting the above equation and inequality into equation (3) leads to

$$\begin{aligned}\Gamma &\ge \sum_m \mathrm{E}[Y \mid A=1,\ M=m]\Pr(M=m \mid A=0) - \mathrm{E}[Y \mid A=1,\ M=1]\\ &= -\{\mathrm{E}[Y \mid A=1,\ M=1] - \mathrm{E}[Y \mid A=1,\ M=0]\}\Pr(M=0 \mid A=0).\end{aligned}$$

A similar calculation gives the upper limit.

Equations (A13) and (A14) hold under Assumption 1 and $E_{01}(1,0) \le E_{01}(1,1)$ holds under Assumption 5 with a = 1, a^* = 0 and m = 1. Substituting (A13) and (A14) into this inequality leads to

$$-(p_1 - p_0)\{\mathrm{E}[Y \mid A=1,\ M=1] - \mathrm{E}[Y \mid A=1,\ M=0]\} \le p_0\gamma_1 + (1-p_0)\gamma_0 = \Gamma.$$

Acknowledgements

This work was supported partially by Grant-in-Aid for Scientific Research (No. 23700344) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Author details

Yasutaka Chiba[1*] and Etsuji Suzuki[2]

1 Division of Biostatistics, Clinical Research Center, Kinki University School of Medicine, Japan

2 Department of Epidemiology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Japan

References

[1] Greenland S, Morgenstern H. Confounding in health research. *Annual Review of Public Health* 2001; 22(1) 189-212.

[2] Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992; 3(2) 143-155.

[3] Cole SR, Hernán MA. Fallibility in estimating direct effects. *International Journal of Epidemiology* 2002; 31(1) 163-165.

[4] VanderWeele T.J. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* 2010; 21(4) 540-551.

[5] Kiely JL. Some conceptual problems in multivariable analyses of perinatal mortality. *Paediatric and Perinatal Epidemiology* 1991; 5(4) 243-257

[6] Kiely JL, Kleinman JC. Birth-weight-adjusted infant mortality in evaluations of perinatal care: towards a useful summary measure. *Statistics in Medicine* 1993; 12(3-4) 377-392.

[7] Kramer MS. Biology vs. methodology in investigating causal pathways for infant mortality. *Paediatric and Perinatal Epidemiology* 2009; 23(5) 414-416.

[8] Hernández-Díaz S, Schisterman EF, Hernán MA. The birth-weight "paradox" uncovered? *American Journal of Epidemiology* 2006; 164(11) 1115-1120.

[9] Schisterman EF, Whitcomb BW, Mumford SL, Platt RW. Z-scores and the birthweight paradox. *Paediatric and Perinatal Epidemiology* 2009; 23(5) 403-413.

[10] Whitcomb BW, Schisterman EF, Perkins NJ, Platt RW. Quantification of collider-stratification bias and the birthweight paradox. *Paediatric and Perinatal Epidemiology* 2009; 23(5) 394-402.

[11] Wilcox AJ, Weinberg CR, Basso O. On the pitfalls of adjusting for gestational age at birth. *American Journal of Epidemiology* 2011; 174(9) 1062-1068.

[12] Judd CM, Kenny DA. Process analysis: estimating mediation in treatment evaluations. *Evaluation Review* 1981; 5(5) 602-619.

[13] Yerushalmy J. The relationship of parents' cigarette smoking to outcome of pregnancy –implications as to the problem of inferring causation from observed associations. *American Journal of Epidemiology* 1971; 93(6) 443-456.

[14] Wilcox AJ. Birthweight and perinatal mortality: the effect of maternal smoking. *American Journal of Epidemiology* 1993; 137(10) 1098-1104.

[15] Platt RW, Joseph KS, Ananth CV, Grondines J, Abrahamowicz M, Kramer MS. A proportional hazards model with time-dependent covariates and time-varying effects for analysis of fetal and infant death. *American Journal of Epidemiology* 2004; 160(4) 199-206.

[16] VanderWeele TJ, Mumford SL, Schisterman EF. Conditioning on intermediates in perinatal epidemiology. *Epidemiology* 2012; 23(1) 1-9.

[17] Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology* 2004; 15(5) 615-625.

[18] VanderWeele TJ, Robins JM. Directed acyclic graphs, sufficient causes and the properties of conditioning on a common effect. *American Journal of Epidemiology* 2007; 166(9) 1096-1104.

[19] Glymour MM, Greenland S. Causal diagrams. In: Rothman KJ, Greenland S, Lash TL (eds.) *Modern Epidemiology* 3rd ed. Philadelphia: Lippincott Williams and Wilkins; 2008. p183-209.

[20] Cole SR, Platt RW, Schisterman EF. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* 2010; 39(2) 417-420.

[21] Basso O, Wilcox AJ. Intersecting birth weight-specific mortality curves: solving the riddle. *American Journal of Epidemiology* 2009; 169(7) 787-797.

[22] Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual Review of Public Health* 2000; 21(1) 121-145.

[23] Hernán MA. A definition of causal effect for epidemiological studies. *Journal of Epidemiology and Community Health* 2004; 58(4) 265-271.

[24] Suzuki E, Yamamoto E, Tsuda T. Identification of operating mediation and mechanism in the sufficient-component cause framework. *European Journal of Epidemiology* 2011; 26(5) 347-357.

[25] Hafeman DM, VanderWeele TJ. Alternative assumptions for the identification of direct and indirect effects. *Epidemiology* 2011; 22(6) 753-764.

[26] Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics* 2002; 58(1) 21-29.

[27] Egleston BL, Cropsey KL, Lazev AB, Heckman CJ. A tutorial on principal stratification-based sensitivity analysis: application to smoking cessation studies. *Clinical Trials* 2010; 7(3) 286-298.

[28] Chiba Y, Taguri M, Uemura Y. On the identification of the survivor average causal effect. *Journal of Biometrics and Biostatistics*, 2011; 2(5) e104.

[29] Hayden D, Pauler DK, Schoenfeld D. An estimator for treatment comparisons amongst survivors in randomized trials. *Biometrics* 2005; 61(1) 305-310.

[30] Chiba Y. Marginal structural models for estimating principal stratum direct effects under the monotonicity assumption. *Biometrical Journal* 2011; 53(6) 1025-1034.

[31] Sjölander A, Humphreys K, Vansteelandt S, Bellocco R, Palmgren J. Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. *Biometrics* 2009; 65(2) 514 -520.

[32] Chiba Y. Bias analysis for the principal stratum direct effect in the presence of confounded intermediate variables. *Journal of Biometrics and Biostatistics* 2010; 1(1) 101.

[33] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables (with discussion). *Journal of the American Statistical Association*, 1996; 91(434) 444-472.

[34] Manski CF. Monotone treatment response. *Econometrica* 1997; 65(6) 1311-1334.

[35] Manski CF, Pepper JV. Monotone instrumental variables: with an application to the returns to schooling. *Econometrica* 2000; 68(4) 997-1010.

[36] Manski CF. *Partial identification of probability distributions*. New York: Springer-Verlag; 2003.

[37] Chiba Y. Causal inference in randomized trials with noncompliance. In: Śmigórski K (ed.) *Health Management – Different Approaches and Solutions*. Rijeka: Intech; 2011. p315-336.

[38] Chiba Y, VanderWeele TJ. A simple method for principal strata effects when the outcome has been truncated due to death. *American Journal of Epidemiology* 2011; 173(7) 745-751.

[39] Pearl J. Direct and indirect effects. In: Breese J, Koller D (eds.) *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 2-5 August 2001. San Francisco: Morgan Kaufmann; 2001. p411-420.

[40] Joffe M, Small D, Hsu CY. Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science* 2007; 22(1) 74-97.

[41] Hafeman DM, Schwartz S. Opening the black box: a motivation for the assessment of mediation. *International Journal of Epidemiology* 2009; 38(4) 838-845.

[42] Petersen ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. *Epidemiology* 2006; 17(3) 276-284.

[43] VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. *Epidemiology* 2009; 20(1) 18-26.

[44] Kaufman S, Kaufman JS, MacLehose RF, Greenland S, Poole C. Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Statistics in Medicine* 2005; 24(11) 1683-1702.

[45] Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* 2010; 25(1) 51-71.

[46] VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis with a dichotomous outcome. *American Journal of Epidemiology* 2010; 172(12) 1339-1348.

[47] Chiba Y. Monte-Carlo sensitivity analysis for controlled direct effects using marginal structural models in the presence of confounded mediators. *Communications in Statistics – Theory and Methods* 2012; 41(10) 1739-1749.

[48] Tchetgen Tchetgen EJ, Shpitser I. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics* (in press).

[49] Chiba Y. Bounds on controlled direct effects under monotonic assumptions about mediators and confounders. *Biometrical Journal* 2010; 52(5) 628-637.

[50] VanderWeele TJ. Simple relations between principal stratification and direct and indirect effects. *Statistics and Probability Letters* 2008; 78(17) 2957-2962.