Bernoulli **22**(1), 2016, 421–443
DOI: 10.3150/14-BEJ664

Equivalence between direct and indirect effects with different sets of intermediate variables and covariates

MANABU KUROKI

Department of Data Science, The Institute of Statistical Mathematics, 10-3, Midori-cho, Tachikawa, Tokyo, 190-8562, Japan. E-mail: mkuroki@ism.ac.jp

This paper deals with the concept of equivalence between direct and indirect effects of a treatment on a response using two sets of intermediate variables and covariates. First, we provide criteria for testing whether two sets of variables can estimate the same direct and indirect effects. Next, based on the proposed criteria, we discuss the variable selection problem from the viewpoint of estimation accuracy of direct and indirect effects, and show that selecting a set of variables that has a direct effect on a response cannot always improve estimation accuracy, which is contrary to the situation found in linear regression models. These results enable us to judge whether different sets of variables can yield the same direct and indirect effects and thus help us select appropriate variables to estimate direct and indirect effects with cost reduction or estimation accuracy.

Keywords: causal effect; equivalence; identification; Markov boundary

1. Introduction

Mediation analysis, which has been discussed in the fields of social science and psychology, is used to evaluate the degree to which intermediate variables (measured temporally between a treatment and a response) mediate the effect of a treatment on a response and has lately attracted considerable attention in practical science. For example, in randomized clinical trials, appropriate intermediate variables are often used as an alternative approach for reducing the cost and duration of the trials, when it is expensive, inconvenient or infeasible within a practical length of time to observe a response. As an example from the field of quality control, intermediate variables are often used to identify the source of a malfunction within a production process before the final quality characteristics (the response) are obtained. In order to choose appropriate intermediate variables to achieve this purpose, it is necessary to clarify how intermediate variables capture the total effect of a treatment on a response.

In general, since intermediate variables do not fully capture the total effect of a treatment on a response (Joffe and Greene [8], Wang and Taylor [32]), it is necessary to decompose the total effect of a treatment on a response into a direct effect not mediated by the intermediate variables and indirect effects mediated through the intermediate variables, and evaluate direct and indirect effects with reasonable estimation accuracy. To formulate the effect decomposition, Pearl [16,17] introduced three distinct causal concepts, which were given as controlled direct effects (CDEs), natural direct effects (NDEs) and natural indirect effects (NIEs), and showed that the total effect can be described by the sum of NDE and NIE. In addition, he proposed the identification conditions for the CDE, NDE and NIE. Imai *et al.* [7], van der Laan and Petersen [30] and other causal researchers discussed alternative identification conditions for the NDE and NIE. The identification problems of the CDE have also been discussed by many researchers, related to the identification conditions for causal effects of joint interventions (Kuroki and Miyakawa [12], Shpitser and Pearl [27], van der Laan and Petersen [30], VanderWeele [29]). Although a great deal of effort has been devoted to establishing identifiability criteria and the methodology for estimating direct and indirect effects, there has been little discussion on whether different sets of intermediate variables and covariates can yield the same estimators when several possible intermediate variables and covariates are available. When the answer is affirmative, the next question would be how to select appropriate variables in order to increase estimation accuracy.

The aim of this paper is to answer the two questions above. First, we provide criteria for testing whether two sets of intermediate variables and covariates can yield the same direct and indirect effects, that is, whether the estimators using one set are guaranteed to yield the same direct and indirect effects as the estimators using the other set. The reason for posing this question is that, given two sets of variables, a researcher may wish to assess, prior to taking any action of experimental studies, whether two candidate sets of variables, differing substantially in dimensionality, cost, data sparseness or measurement error can yield the same direct and indirect effects. Next, based on the proposed criteria, we discuss the variable selection problem from the viewpoint of the estimation accuracy of the NDE and NIE for discrete cases, and show that selecting a set of variables that has a direct effect on a response cannot always improve the estimation accuracy even in ideal experimental studies, which is contrary to the situation found in linear regression models (e.g., Kuroki and Cai [10], Kuroki and Miyakawa [13]). These results help us select appropriate set of variables to reduce cost without amplifying the bias related to the direct and indirect effects.

This paper is organized as follows. Section 2 gives some preliminary considerations that will be used throughout the paper. In Section 3, we introduce the concept of equivalence in which two sets of variables provide the same (asymptotic) bias for the estimates of direct and indirect effects. Then, we provide sufficient conditions for equivalence between two sets of variables. Section 4 discusses the variable selection problem from the viewpoint of the estimation accuracy. Simulation experiments verifying our results are presented in Section 5. Finally, Section 6 concludes this paper.

2. Preliminaries

2.1. Potential response approach

In order to discuss our problem, we use the potential response approach (Pearl [17], Rubin [24, 25]). Let X, S and Y be a treatment, an intermediate variable and a response, respectively. Letting D_X, D_S and D_Y be the domains of X, S and Y respectively, we let x, s and y represent the values taken by the variables X, S and Y, respectively ($x \in D_X, s \in D_S, y \in D_Y$). Similar notation is used for other variables, domains and values. In addition, we use $Y_x(i) = y$ and $Y_{x,s}(i) = y$, which are called potential responses, to denote respectively the counterfactual sentences "Y

would have the value y, had X been x for the ith subject" and "Y would have the value y, had X and S been x and s for the ith subject, respectively". Similar notation is used for other potential responses.

In this paper, we assume the stable unit treatment value assumption (SUTVA) which consists of the "no interference between units" assumption and the "consistency" assumption. The "no interference between units" assumption means that $Y_x(i)$ and $Y_{x,s}(i)$ ($x \in D_X, s \in D_S$) for the ith subject is not dependent on the treatment or the intermediate variable received by other subjects (Rubin [26]). When n subjects in the study are considered random samples from the population under consideration, since $Y_x(i)$ and $Y_{x,s}(i)$ can be referred to as random variables Y_x and $Y_{x,s}$ respectively, probabilities of potential responses can be defined as $\mathrm{pr}(Y_x = y) \stackrel{\triangle}{=} \mathrm{pr}(y_x)$ and $\mathrm{pr}(Y_{x,s} = y) \stackrel{\triangle}{=} \mathrm{pr}(y_{x,s})$, where $\mathrm{pr}(X = x)$ indicates a marginal probability of $X = x$. Similar notation is used for other marginal probabilities. In addition, $Y_x(i)$ is observed if the ith subject has received $X = x$, and $Y_{x,s}(i)$ is observed if the ith subject has received both $X = x$ and $S = s$. This is called the consistency (Pearl [17], Robins [19,20], Rubin [26]), which is another part of SUTVA and is formulated as "$X = x \Rightarrow Y_x = Y$" and "$X = x$ and $S = s \Rightarrow Y_{x,s} = Y$". The consistency assumption, for example, "$X = x \Rightarrow Y_x = Y$", means that the values of Y which would have been observed if X had been set to what it in fact was are equal to the values of Y which were in fact observed, that is, if the actual value of X turns out to be x, then the value that Y would take on if X were x is consistent with the actual value of Y for every subject.

When a randomized experiment is conducted, since X is independent of Y_x for any $x \in D_X$, which is denoted as $X \perp\!\!\!\perp Y_x$ for any $x \in D_X$, we have $\mathrm{pr}(y_x) = \mathrm{pr}(y|x)$ from the consistency assumption, where $\mathrm{pr}(y|x)$ is a conditional probability of $Y = y$ given $X = x$. Similar notation is used for other conditional probabilities. On the other hand, when a randomized experiment is difficult to conduct and only observational data is available, if there exists such a set $\mathbf{Z}$ of observed covariates that X is conditionally independent of Y_x given $\mathbf{Z}$ for any $x \in D_X$, which is denoted as $X \perp\!\!\!\perp Y_x|\mathbf{Z}$ for any $x \in D_X$, and $\mathrm{pr}(x|\mathbf{z}) > 0$ for any x and $\mathbf{z}$, $\mathrm{pr}(y_x)$ is identifiable by using $\mathbf{Z}$ and is given by $E_z\{\mathrm{pr}(y|x, \mathbf{Z})\}$ (Rosenbaum and Rubin [23]). Here, "identifiability" means that the causal quantities such as $\mathrm{pr}(y_x)$ can be estimated consistently from a joint distribution of observed variables and $E_z\{\mathrm{pr}(y|x, \mathbf{Z})\}$ is the expectation of $\mathrm{pr}(y|x, \mathbf{Z})$ regarding $\mathbf{Z}$.

2.2. Direct and indirect effects

Pearl [16,17] introduced three different concepts of causal quantities, which are "controlled direct effect (CDE)", "natural direct effect (NDE)" and "natural indirect effect (NIE)", and showed that "total effect (TE)" can be described by the sum of the NDE and NIE. For $x, x' \in D_X$ and $s \in D_S$, the CDE of X on Y comparing $X = x$ and $X = x'$ and setting an intermediate variable S to some value s measures the effect of X on Y not mediated through S, that is, the causal effect of X on Y after intervening to fix an intermediate variable S to some value s. Then, the CDE is defined by $\mathrm{CDE}_y^s(x, x') = \mathrm{pr}(y_{x,s}) - \mathrm{pr}(y_{x',s})$. The NDE, which Robins and Greenland [21] called a "pure" direct effect, is different from the CDE in the sense that an intermediate variable S is set to the level $S_{x'}$, which is the level it would have naturally adopted under $X = x'$. Thus, the NDE is defined as $\mathrm{NDE}_y^S(x, x') = \mathrm{pr}(y_{x,S_{x'}}) - \mathrm{pr}(y_{x',S_{x'}})$. Similarly, the NIE, which Hafeman and Schwartz [6] called a "total" indirect effect, is defined by $\mathrm{NIE}_y^S(x, x') = \mathrm{pr}(y_{x,S_x}) - \mathrm{pr}(y_{x,S_{x'}})$

in this paper, which compares the effect of an intermediate variable S at levels S_x and $S_{x'}$ on the response when X is set to x. The TE of X on Y comparing $X = x$ and $X = x'$ measures the overall effect of X on Y. According to the composition property, that is, $Y_x = Y_{x,S_x}$ for $X = x$ (Pearl [17]), the TE of X on Y, $\mathrm{TE}_y(x, x') = \mathrm{pr}(y_x) - \mathrm{pr}(y_{x'})$ can be decomposed as the sum of the NDE and NIE because we have

$$\begin{aligned}\mathrm{TE}_y\big(x, x'\big) &= \mathrm{pr}(y_x) - \mathrm{pr}(y_{x,S_{x'}}) + \mathrm{pr}(y_{x,S_{x'}}) - \mathrm{pr}(y_{x'}) \\ &= \mathrm{pr}(y_{x,S_x}) - \mathrm{pr}(y_{x,S_{x'}}) + \mathrm{pr}(y_{x,S_{x'}}) - \mathrm{pr}(y_{x',S_{x'}}) = \mathrm{NIE}_y^S\big(x, x'\big) + \mathrm{NDE}_y^S\big(x, x'\big).\end{aligned}$$

In this paper, we assume that:

(a) a set of covariates $\mathbf{Z}$ satisfies both $\mathbf{S} \perp\!\!\!\perp Y_{x,s} | \{X\} \cup \mathbf{Z}$ and $Y_{x,s} \perp\!\!\!\perp \mathbf{S}_{x'} | \mathbf{Z}$ for $x, x' \in D_X$ and any $\mathbf{s} \in D_S$, and
(b) a randomized experiment for the treatment X is conducted, that is, $X \perp\!\!\!\perp \{Y_{x,s}\} \cup \mathbf{S}_{x'} \cup \mathbf{Z}$ for $x, x' \in D_X$ and any $\mathbf{s} \in D_S$.

This situation, which is discussed by many researchers (Cai *et al.* [2], Kaufman *et al.* [9]), can be described by the directed acyclic graph shown in Figure 1. For the graph terminology used in this paper, see Pearl [17].

In Figure 1, the directed arrow from X to Y indicates that X could have a direct effect on Y without being mediated by S. In addition, the absence of an arrow pointing from S to X indicates that S does not cause X, and the directed path from X to Y through S indicates that X could also have an effect on Y mediated by S. Furthermore, directed arrows from Z to both S and Y mean that Z could have effects on both S and Y without being mediated by other variables in the graph. When a directed acyclic graph such as Figure 1 indicates the data generating process, conditional independence relationships between variables can be read off from the graph through the d-separation criterion, that is, if $\mathbf{C}$ d-separates $\mathbf{A}$ from $\mathbf{B}$ then $\mathbf{A}$ is conditionally independent of $\mathbf{B}$ given $\mathbf{C}$ (Pearl [15]). For example, since an empty set d-separates X from $\mathbf{Z}$ in Figure 1, X is independent of $\mathbf{Z}$. For details on d-separation criterion, see Pearl [15]. Then, for example, the graph-based causal inference and the potential response approach can be connected by the following rules. For details, refer to Pearl [17].

Exclusion restrictions: For every variable Y having parents $\mathrm{PA}(Y)$ and for every set of variables $\mathbf{S}$ disjoint of $\mathrm{PA}(Y)$, we have $Y_{\mathrm{pa}(Y)} = Y_{\mathrm{pa}(Y),s}$.

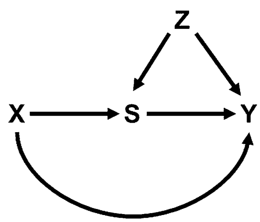

Figure 1. Problem description by a directed acyclic graph for X, S, Y, and Z representing a treatment, an intermediate variable, a response, and a covariate, respectively.

Independence restrictions: If $Z_1, \dots, Z_k$ is any set of variables not connected to Y via dashed arcs, we have $Y_{\mathrm{pa}(Y)} \perp\!\!\!\perp \{Z_{1,\mathrm{pa}(Z_1)}, \dots, Z_{k,\mathrm{pa}(Z_k)}\}$.

Let $\mathrm{CDE}_y^s(x, x'; \mathbf{Z})$, $\mathrm{NDE}_y^S(x, x'; \mathbf{Z})$ and $\mathrm{NIE}_y^S(x, x'; \mathbf{Z})$ be the CDE, NDE and NIE when a set of covariates $\mathbf{Z}$ is used respectively. Then, the CDE, NDE, NIE and TE are identifiable through the observation of X, Y, $\mathbf{S}$ and $\mathbf{Z}$ and are given by

$$\left.\begin{aligned}
\mathrm{CDE}_y^s\big(x, x'; \mathbf{Z}\big) &= \sum_z \big\{\mathrm{pr}(y|x, \mathbf{s}, \mathbf{z}) - \mathrm{pr}\big(y|x', \mathbf{s}, \mathbf{z}\big)\big\}\, \mathrm{pr}(\mathbf{z}),\\
\mathrm{NDE}_y^S\big(x, x'; \mathbf{Z}\big) &= \sum_{s,z} \big\{\mathrm{pr}(y|x, \mathbf{s}, \mathbf{z}) - \mathrm{pr}\big(y|x', \mathbf{s}, \mathbf{z}\big)\big\}\, \mathrm{pr}\big(\mathbf{s}|x', \mathbf{z}\big)\, \mathrm{pr}(\mathbf{z}),\\
\mathrm{NIE}_y^S\big(x, x'; \mathbf{Z}\big) &= \sum_{s,z} \mathrm{pr}(y|x, \mathbf{s}, \mathbf{z}) \big\{\mathrm{pr}(\mathbf{s}|x, \mathbf{z}) - \mathrm{pr}\big(\mathbf{s}|x', \mathbf{z}\big)\big\}\, \mathrm{pr}(\mathbf{z}),\\
\mathrm{TE}_y\big(x, x'\big) &= \mathrm{pr}(y|x) - \mathrm{pr}\big(y|x'\big) = \mathrm{NDE}_y^S\big(x, x'; \mathbf{Z}\big) + \mathrm{NIE}_y^S\big(x, x'; \mathbf{Z}\big),
\end{aligned}\right\}$$

respectively. Especially, since we have the condition $X \perp\!\!\!\perp \mathbf{Z}$, the NDE and NIE can be rewritten as

$$\left.\begin{aligned}
\mathrm{NDE}_y^S\big(x, x'; \mathbf{Z}\big) &= \sum_{s,z} \big\{\mathrm{pr}(y|x, \mathbf{s}, \mathbf{z}) - \mathrm{pr}\big(y|x', \mathbf{s}, \mathbf{z}\big)\big\}\, \mathrm{pr}\big(\mathbf{s}, \mathbf{z}|x'\big),\\
\mathrm{NIE}_y^S\big(x, x'; \mathbf{Z}\big) &= \sum_{s,z} \mathrm{pr}(y|x, \mathbf{s}, \mathbf{z}) \big\{\mathrm{pr}(\mathbf{s}, \mathbf{z}|x) - \mathrm{pr}\big(\mathbf{s}, \mathbf{z}|x'\big)\big\},
\end{aligned}\right\} \tag{1}$$

respectively. In this paper, equations (1) form the basis of our discussion. Here, it is noted that summation is replaced by integration whenever the variables are continuous. The discussion in Section 3 is based on nonparametric models. However, in Sections 4 and 5, it is assumed that the variables of interests follow a multinomial distribution.

3. Equivalence between variables

3.1. Motivation and definition

We illustrate our motivation using a case study from quality control (Technometrics Research Group [28]). The IC (Integrated Circuit) manufacturing line was constructed by hundred elementary processes which were connected in series. Technometrics Research Group [28] was interested in how the gate oxide thickness (X) in the process of the gate oxide formation has a direct effect on the threshold voltage (Y) not through the heat treatment process. They considered several settings in this case study. Initially, they assumed the causal chain $X \to S_1 \to S_2 \to Y$ based on the IC manufacturing line and measured the resistances of the P-type channel (S_1) and a certain characteristic (S_2) in order to monitor the effect of the heat treatment process on Y. However, since it was known that X had an effect on Y but we did not know how large it was, Technometrics Research Group [28] considered the directed acyclic graph corresponding to this manufacturing line shown in Figure 2. Then, they applied the linear regression analysis of Y on X, S_1 and S_2 to observed data with sample size $n = 29$, and found that the regression coefficient of X was not statistically significant, which indicated that the gate oxide thickness (X) did not

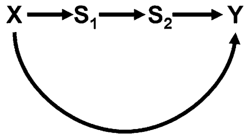

Figure 2. The simple situation of the IC manufacturing line.

have a significant direct effect on the threshold voltage (Y). Here, confounders may exist between $\{S_1, S_2\}$ and Y but they were ignored in Technometrics Research Group [28]. Therefore, we assume that no confounders exist in this case study.

In this paper, we will show that S_1, S_2 and $\{S_1, S_2\}$ can provide the same (asymptotic) estimators of the direct effect (and indirect effects) in the situation shown in Figure 2. That is, when Figure 2 reflects the IC manufacturing line, according to our results, it is not necessary to observe both S_1 and S_2 but either of them is enough in order to estimate the direct effect of X on Y. Although some of the proposed conditions are not described based on the terms of graphical causal inference (Pearl [17]), if we know that the IC manufacturing line can be described by Figure 2 before actual observation, we can provide such judgment from the graph structure, through the relationships between the d-separation criterion and statistical independencies. As a result, it is expected to reduce cost and save time. For example, when the correlation matrix shown in Table 1 is assumed to be derived according to Figure 2, the direct effect of X on Y are estimated by $\hat{\beta}_{yx.s} = -0.063$ whichever we use $S = S_1$, S_2 or $\{S_1, S_2\}$, where $\hat{\beta}_{yx.s}$ is an ordinary least square estimator of the regression coefficient $\beta_{yx.s}$ of X in the linear regression model of Y on X and S. Similar notation is used for other regression coefficients.

In Figure 2, since both $Y \perp\!\!\!\perp S_1 | \{X, S_2\}$ and $S_2 \perp\!\!\!\perp X | S_1$ hold, whichever we use S_1, S_2 or $\{S_1, S_2\}$, the NDE and NIE can be provided by $\mathrm{NDE}_y^{S_i}(x, x'; \phi) = \mathrm{NDE}_y^{S_1,S_2}(x, x'; \phi)$ and $\mathrm{NIE}_y^{S_i}(x, x'; \phi) = \mathrm{NIE}_y^{S_1,S_2}(x, x'; \phi)$ respectively ($i = 1, 2$) from the proposed conditions, which implies that the statistics of NDE and NIE using S_i ($i = 1, 2$) can estimate the same NDE and NIE as those using both S_1 and S_2. According to this consideration, we introduce the concept of equivalence between two sets of variables in the sense that the same causal quantity can be estimated whichever set of variables is used, where we say "**A** and **B** are different sets" for two sets **A** and **B** of variables when $\mathbf{A} \neq \mathbf{B}$ holds.

Definition 1 (Equivalence given x ***and*** x'***).*** *For two sets of variables* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *and given values* x *and* x' *of interest* $(x, x' \in D_X)$, *they are equivalent to each other given* x *and* x' *relative*

Table 1. Correlation matrix based on Figure 2

	X	S_1	S_2	Y
X	1.000	−0.428	0.088	−0.132
S_1	−0.428	1.000	−0.206	0.188
S_2	0.088	−0.206	1.000	−0.787
Y	−0.132	0.188	−0.787	1.000

to (X, Y), *if the following equality holds for any* y;

$$\sum_{t_1} \mathrm{pr}(y|x, \mathbf{t}_1)\, \mathrm{pr}\big(\mathbf{t}_1|x'\big) = \sum_{t_2} \mathrm{pr}(y|x, \mathbf{t}_2)\, \mathrm{pr}\big(\mathbf{t}_2|x'\big), \tag{2}$$

where the LHS (*RHS*) *of equation* (2) *is replaced by* $\mathrm{pr}(y|x)$ *when* $\mathbf{T}_1$ $(\mathbf{T}_2)$ *is an empty set.*

Trivially, if $\mathbf{T}_1$ is the same as $\mathbf{T}_2$ then $\mathbf{T}_1$ and $\mathbf{T}_2$ are equivalent to each other given x and x'. In addition, if $x = x'$ holds, then $\mathbf{T}_1$ and $\mathbf{T}_2$ are always equivalent to each other given x and x'. Thus, we do not discuss these cases. In Figure 1, Z is equivalent to an empty set given x and x' but not to a set including S in general. On the other hand, in Figure 2, S_1, S_2 and $\{S_1, S_2\}$ are equivalent to each other given x and x' (S_i $(i = 1, 2)$ and $\{S_1, S_2\}$ are different sets in the sense that one of the elements in $\{S_1, S_2\}$ is not included in $\{S_i\}$).

If the same causal quantities can be estimated whichever a set of variables is used, then we can choose better a set of variables in terms of estimation accuracy, dimensionality of intermediate variables, data-sparseness, or cost reduction. In that sense, the concept of equivalence plays an important role in the evaluation of causal quantities such as total effects, direct and indirect effects.

When we consider Definition 1 for any x', we have

$$\sum_{t_1} \mathrm{pr}(y|x, \mathbf{t}_1)\, \mathrm{pr}(\mathbf{t}_1) = \sum_{t_2} \mathrm{pr}(y|x, \mathbf{t}_2)\, \mathrm{pr}(\mathbf{t}_2)$$

from equation (2). Thus, Definition 1 can be regarded as the weaker version of the definition of equivalence proposed by Pearl [18] in the sense that the latter is based on the whole population but the former is based on the subpopulation $X = x'$. For this reason, equivalence given x and x' is called weak equivalence throughout this paper. On the other hand, when X is a dichotomous variable, for a non-empty set $\mathbf{T}$, we have

$$\sum_{t} \mathrm{pr}(y|x, \mathbf{t})\, \mathrm{pr}\big(\mathbf{t}|x'\big) = \frac{\sum_t \mathrm{pr}(y|x, \mathbf{t})\, \mathrm{pr}(\mathbf{t}) - \mathrm{pr}(x, y)}{\mathrm{pr}(x')}. \tag{3}$$

Thus, Definition 1 is essentially the same as the concept of the equivalence proposed by Pearl [18] in this case.

One important application of the equivalence is the propensity score using intermediate variables and covariates, that is, $0 < PS = \mathrm{pr}(x|\mathbf{z}, \mathbf{s}) < 1$ when X is a dichotomous variable ($D_X = \{x, x'\}$). When $\mathbf{Z}$ and $\mathbf{S}$ satisfy conditions (a) and (b) in Section 2.2, since we have $X \perp\!\!\!\perp \mathbf{S} \cup \mathbf{Z}|PS$ by tracing the proof of Theorem 2 in Rosenbaum and Rubin [23] and $Y \perp\!\!\!\perp PS|\mathbf{S} \cup \mathbf{Z} \cup \{X\}$ because $\mathrm{pr}(y|x, ps, \mathbf{s}, \mathbf{z}) = \mathrm{pr}(y|x, \mathbf{s}, \mathbf{z})$ and $\mathrm{pr}(y|x', ps, \mathbf{s}, \mathbf{z}) = \mathrm{pr}(y|x', \mathbf{s}, \mathbf{z})$ hold, the propensity score is weakly equivalent to $\mathbf{S} \cup \mathbf{Z}$ regarding the NDE and NIE. Thus, when we estimate direct and indirect effects, the propensity score can be used for reducing the dimensionality of a large set of variables to unity.

3.2. Sufficient conditions for weak equivalence

In this section, we provide some sufficient conditions for weak equivalence.

Theorem 1. *For x and x', if* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *relative to* (X, Y) *satisfies one of the following two conditions then* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *are weakly equivalent to each other*: (i) $X \perp\!\!\!\perp (\mathbf{T}_2 \setminus \mathbf{T}_1)|\mathbf{T}_1$ *and* $Y \perp\!\!\!\perp (\mathbf{T}_1 \setminus \mathbf{T}_2)|\{X\} \cup \mathbf{T}_2$ *and* (ii) $X \perp\!\!\!\perp (\mathbf{T}_1 \setminus \mathbf{T}_2)|\mathbf{T}_2$ *and* $Y \perp\!\!\!\perp (\mathbf{T}_2 \setminus \mathbf{T}_1)|\{X\} \cup \mathbf{T}_1$.

Proof. For condition (i), we have

$$\sum_{t_1} \mathrm{pr}(y|x, \mathbf{t}_1)\,\mathrm{pr}(\mathbf{t}_1|x') = \sum_{t_1 \cup t_2} \mathrm{pr}(y|x, \mathbf{t}_1 \cup \mathbf{t}_2)\,\mathrm{pr}(\mathbf{t}_2 \setminus \mathbf{t}_1|x, \mathbf{t}_1)\,\mathrm{pr}(\mathbf{t}_1|x')$$

$$= \sum_{t_1 \cup t_2} \mathrm{pr}(y|x, \mathbf{t}_2)\,\mathrm{pr}(\mathbf{t}_2 \setminus \mathbf{t}_1|x', \mathbf{t}_1)\,\mathrm{pr}(\mathbf{t}_1|x') = \sum_{t_2} \mathrm{pr}(y|x, \mathbf{t}_2)\,\mathrm{pr}(\mathbf{t}_2|x').$$

Condition (ii) can also be achieved by the similar way. □

As seen from the proof of Theorem 1, if the conditions of Theorem 1 hold, then $\mathbf{T}_1 \cup \mathbf{T}_2$ is also weakly equivalent to $\mathbf{T}_i$ $(i = 1, 2)$. In addition, for example, when we have $\mathbf{T}_1 \subset \mathbf{T}_2$, if either $X \perp\!\!\!\perp \mathbf{T}_2 \setminus \mathbf{T}_1|\mathbf{T}_1$ or $Y \perp\!\!\!\perp \mathbf{T}_2 \setminus \mathbf{T}_1|\{X\} \cup \mathbf{T}_1$ holds, then $\mathbf{T}_1$ and $\mathbf{T}_2$ are weakly equivalent to each other by tracing the proof of Theorem 1.

The intuition behind Theorem 1 is easy to understand through the collapsibility conditions in linear regression models: for two linear regression models, the full model of Y on X, T_1 and T_2, that is, $Y = \beta_{y.xt_1t_2} + \beta_{yx.t_1t_2}X + \beta_{yt_1.xt_2}T_1 + \beta_{yt_2.xt_1}T_2 + \varepsilon_{y.xt_1t_2}$ and the reduced model of Y on X and T_1, that is, $Y = \beta_{y.xt_1} + \beta_{yx.t_1}X + \beta_{yt_1.x}T_1 + \varepsilon_{y.xt_1}$ with Gaussian errors $\varepsilon_{y.xt_1t_2}$ and $\varepsilon_{y.xt_1}$, we will say that T_2 is collapsible with respect to (X, Y) relationship when $\beta_{yx.t_1t_2} = \beta_{yx.t_1}$ holds. It is well known that T_2 is collapsible with respect to (X, Y) relationship when $X \perp\!\!\!\perp T_2|T_1$ or $Y \perp\!\!\!\perp T_2|\{X, T_1\}$ holds (e.g., Clogg *et al.* [3], Kuroki and Cai [10], Kuroki and Miyakawa [13], Wermuth [33]). Different from the collapsibility conditions that focus on the dimension reduction in the sense whether the regression coefficient of X is unchanged by removing T_2 from the full model, equivalence focuses on whether two regression models of Y on X and T_1 and Y on X and T_2 (asymptotically) provide the same estimates for the regression coefficients of X.

Here, it is noted that the conditions offered by Theorem 1 do not characterize all weak equivalence pairs. For example, when we consider the NDE and NIE of X on Y through $\{S_1, S_4\}$ in Figure 3, although another set $\{S_2, S_3\}$ can provide the same NDE and NIE of X on Y through $\{S_1, S_4\}$, thus, they must be weakly equivalent to each other, neither (i) or (ii) holds in this case.

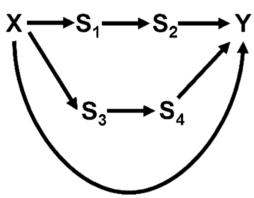

Figure 3. Since $\{S_1, S_3\}$ and $\{S_2, S_4\}$ satisfy conditions in Theorem 1, they are weakly equivalent to each other. Although $\{S_1, S_4\}$ and $\{S_2, S_3\}$ are also weakly equivalent to each other, they do not satisfy conditions in Theorem 1.

Theorem 2. *Letting* $\mathbf{T}_i^m$ *be a subset of* $\mathbf{T}_i$ *satisfying* $X \perp\!\!\!\perp (\mathbf{T}_i \setminus \mathbf{T}_i^m)|\mathbf{T}_i^m$ $(i = 1, 2)$, *if* $\mathbf{T}_1^m = \mathbf{T}_2^m$ *holds then* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *are weakly equivalent to each other.*

The proof is obvious: since $\mathbf{T}_1^m = \mathbf{T}_2^m = \mathbf{T}^m$ is a subset of both $\mathbf{T}_1$ and $\mathbf{T}_2$ from the assumption, $\mathbf{T}^m$ is weakly equivalent to both $\mathbf{T}_1$ and $\mathbf{T}_2$ as seen from the proof of Theorem 1. Thus, $\mathbf{T}_1$ and $\mathbf{T}_2$ are also weakly equivalent to each other. Theorem 2 states that if two sets include the same set of variables which make a treatment and the remaining variables conditionally independent then they are weakly equivalent to each other.

A subset $\mathbf{T}^m \subset \mathbf{T}$ of variables satisfying $X \perp\!\!\!\perp (\mathbf{T} \setminus \mathbf{T}^m)|\mathbf{T}^m$ is often called a (Markov) blanket of X relative to $\mathbf{T}$, and the minimal Markov blanket is called a Markov boundary in the context of graphical models (Pearl [15]).

Theorem 3. *If* $\mathbf{U}$ *is a Markov boundary of* Y *relative to* $\mathbf{T}_1 \cup \mathbf{T}_2 \cup \{X\}$ *satisfying* $X \perp\!\!\!\perp ((\mathbf{U} \setminus \{X\}) \cap (\mathbf{T}_{3-i} \setminus \mathbf{T}_i))|\mathbf{T}_i$ $(i = 1, 2)$, *then* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *are weakly equivalent to each other.*

Proof. We have

$$\begin{aligned}
\sum_{t_i} \mathrm{pr}(y|x, \mathbf{t}_i)\,\mathrm{pr}\big(\mathbf{t}_i|x'\big) &= \sum_{t_1 \cup t_2} \mathrm{pr}(y|x, \mathbf{t}_1 \cup \mathbf{t}_2)\,\mathrm{pr}(\mathbf{t}_{3-i} \setminus \mathbf{t}_i|x, \mathbf{t}_i)\,\mathrm{pr}\big(\mathbf{t}_i|x'\big) \\
&= \sum_{t_i, u \setminus \{x\}} \mathrm{pr}\big(y|x, \mathbf{u} \setminus \{x\}\big)\,\mathrm{pr}\big(\big(\mathbf{u} \setminus \{x\}\big) \cap (\mathbf{t}_{3-i} \setminus \mathbf{t}_i)|\mathbf{t}_i, x\big)\,\mathrm{pr}\big(\mathbf{t}_i|x'\big) \\
&= \sum_{t_i, u \setminus \{x\}} \mathrm{pr}\big(y|x, \mathbf{u} \setminus \{x\}\big)\,\mathrm{pr}\big(\big(\mathbf{u} \setminus \{x\}\big) \cap (\mathbf{t}_{3-i} \setminus \mathbf{t}_i), \mathbf{t}_i|x'\big) \\
&= \sum_{u \setminus \{x\}} \mathrm{pr}\big(y|x, \mathbf{u} \setminus \{x\}\big)\,\mathrm{pr}\big(\mathbf{u} \setminus \{x\}|x'\big),
\end{aligned}$$

thus, the theorem is proved. □

Theorem 3 is different from Theorem 2 in the sense that Theorem 3 is based on the Markov boundary of the response but not that of the treatment. Intuitively, when there is no confounder between S and Y, Theorem 2 is used to select a set of variables which are direct effects (children of the treatment) or "more close to" the treatment from a given set of variables. On the other hand, Theorem 3 selects a set of variables which are direct causes (parents of the response) or "more close to" the response from a given set of variables. In addition, for example, when we have $\mathbf{T}_1 \subset \mathbf{T}_2$, if $\mathbf{U}$ is a Markov boundary of the response Y relative to $\mathbf{T}_2 \cup \{X\}$ satisfying $X \perp\!\!\!\perp ((\mathbf{U} \setminus \{X\}) \cap (\mathbf{T}_2 \setminus \mathbf{T}_1))|\mathbf{T}_1$, then $\mathbf{T}_1$ and $\mathbf{T}_2$ are weakly equivalent to each other by tracing the proof of Theorem 3. For example, in Figure 3, although $\{S_1, S_4\}$ and $\{S_2, S_3\}$ are also weakly equivalent to each other, they do not satisfy conditions in Theorem 1 or Theorem 2 (because $\{S_1, S_4\} \cap \{S_2, S_3\} = \phi$) but satisfy conditions in Theorem 3.

Finally, as an example that the proposed sufficient conditions in this section do not hold but two sets are weakly equivalent to each other, we consider a joint probability shown in Table 2. Letting $T_1 = \{Z\}$ and $T_2 = \{S\}$, since $\{S, Z\}$ is a Markov boundary of Y relative to $\{X, S, Z\}$

Table 2. A joint probability $\mathrm{pr}(x, y, s, z)$ that the proposed sufficient conditions do not hold but S and Z are weakly equivalent to each other

		x_1		x_0	
		z_1	z_0	z_1	z_0
y_1	s_1	0.038	0.073	0.024	0.021
	s_0	0.067	0.234	0.025	0.114
y_0	s_1	0.004	0.073	0.001	0.018
	s_0	0.100	0.114	0.040	0.054

but neither $X \perp\!\!\!\perp Z|S$ or $X \perp\!\!\!\perp S|Z$ hold, the sufficient conditions of Theorem 3 do not hold. In addition, since $T_1 \cap T_2$ is an empty set and we have $X \perp\!\!\!\perp Z$ but $X \not\perp\!\!\!\perp S$, the sufficient conditions of Theorem 2 do not hold. Furthermore, we have $Y \perp\!\!\!\perp X|\{S, Z\}$ but neither (i) $X \perp\!\!\!\perp S|Z$ and $Y \perp\!\!\!\perp Z|\{X, S\}$ or (ii) $X \perp\!\!\!\perp Z|S$ and $Y \perp\!\!\!\perp S|\{X, Z\}$ hold, thus the sufficient conditions of Theorem 1 do not hold. However, we know that Z is weakly equivalent to S because we have $\sum_s \mathrm{pr}(y|x_1, s)\,\mathrm{pr}(s|x_0) = \sum_z \mathrm{pr}(y|x_1, z)\,\mathrm{pr}(z|x_0) = 0.586$.

This example shows that other sufficient conditions could be derived through a precise parameter tuning.

4. Variable selection for estimating the NDE and NIE for discrete variables

4.1. Motivation

Technometrics Research Group [28] was interested in the evaluation of the direct effect of the gate oxide thickness (X) on the threshold voltage (Y) not through the heat treatment process in the case study of Section 3.1. When we assume that (X, S_1, S_2, Y) follows the multivariate normal distribution based on Technometrics Research Group [28], according to Kuroki and Cai [10], we can read off from Figure 2 that S_2 can (asymptotically) provide better estimation accuracy of the direct effect of X on Y because both $X \perp\!\!\!\perp S_2|S_1$ and $Y \perp\!\!\!\perp S_1|\{S_2, X\}$ hold (intuitively, S_2 is a direct cause of Y). Actually, we have $\sqrt{\mathrm{a.var}(\hat{\beta}_{yx.s_1s_2})} = 0.1261$, $\sqrt{\mathrm{a.var}(\hat{\beta}_{yx.s_1})} = 0.2015$ and $\sqrt{\mathrm{a.var}(\hat{\beta}_{yx.s_2})} = 0.1144$ from Table 1. Here, "a.var $(\cdot)$" is the asymptotic variance of the estimator in parentheses. That is, based on Figure 2, we judge that S_2 should be used if one wish to estimate the direct effect of X on Y with better (asymptotically) estimation accuracy. However, such a result may not hold for discrete cases. Therefore, we consider the variable selection in discrete cases in the next section.

Table 3. Data layout in stratum U

	y_1	y_2	
x_1	$n_{x_1,y_1,u}$	$n_{x_1,y_2,u}$	$n_{x_1,u}$
x_2	$n_{x_2,y_1,u}$	$n_{x_2,y_2,u}$	$n_{x_2,u}$
	$n_{y_1,u}$	$n_{y_2,u}$	n_u

4.2. Variance estimators for discrete variables

In this section, to propose variance estimators for the NDE and NIE presented as equations (1) when both X and Y are dichotomous variables, we consider a contingency table shown in Table 3. When $\mathbf{S}$ and $\mathbf{Z}$ are sets of discrete intermediate variables and covariates satisfying conditions (a) and (b) in Section 2.2 respectively, Table 3 shows the observed subjects in stratum $U = u$, for a non-empty set $U = \mathbf{S} \cup \mathbf{Z}$. We assume that $n_{x_1,y_1,u}$ subjects develop the disease ($Y = y_1$) in the treated group ($X = x_1$) in stratum $U = u$. Similar notation is used for other frequencies. In this paper, we assume that $n_{x_i,y_j,u}$ $(i, j = 1, 2; u = u_1, \ldots, u_p)$ follow the multinomial distribution $\mathrm{MN}(n_{x_i}, \{\mathrm{pr}(y_j, u|x_i) | j = 1, 2; u = u_1, \ldots, u_p\})$ for $i = 1, 2$, where $n_x = \sum_{u,y} n_{x,y,u}$ $(x \in \{x_1, x_2\})$.

Under this situation, $\mathrm{pr}(y, u|x)$ is estimated by $n_{x,y,u}/n_x$ $(x \in \{x_1, x_2\}, y \in \{y_1, y_2\}, u \in \{u_1, \ldots, u_p\})$. Then, the variances of $\widehat{\mathrm{NDE}}^S_y(x_1, x_2; \mathbf{Z})$ and $\widehat{\mathrm{NIE}}^S_y(x_1, x_2; \mathbf{Z})$ are given by

$$\begin{aligned}
&\mathrm{var}\{\widehat{\mathrm{NDE}}^S_y(x_1, x_2; \mathbf{Z})\} \\
&= \sum_u \{\mathrm{pr}(y|x_1, u) - \mathrm{pr}(y|x_2, u)\}^2 \frac{\mathrm{pr}(u|x_2)}{n_{x_2}} - \frac{\mathrm{NDE}^{S2}(x_1, x_2; \mathbf{Z})}{n_{x_2}} \\
&\quad + \sum_u \frac{\mathrm{pr}(y|x_1, u)(1 - \mathrm{pr}(y|x_1, u))}{n^2_{x_2}} E\left(\frac{n^2_{x_2,u}}{n_{x_1,u}}\right) + \sum_u \frac{\mathrm{pr}(y|x_2, u)(1 - \mathrm{pr}(y|x_2, u))}{n_{x_2}} \mathrm{pr}(u|x_2),
\end{aligned} \tag{4}$$

$$\begin{aligned}
&\mathrm{var}\{\widehat{\mathrm{NIE}}^S_y(x_1, x_2; \mathbf{Z})\} \\
&= \sum_u \frac{\mathrm{pr}(y|x_1, u)(1 - \mathrm{pr}(y|x_1, u))}{n^2_{x_2}} E\left(\frac{n^2_{x_2,u}}{n_{x_1,u}}\right) \\
&\quad - 2 \sum_u \mathrm{pr}(y|x_1, u)\big(1 - \mathrm{pr}(y|x_1, u)\big) \frac{\mathrm{pr}(u|x_2)}{n_{x_1}} + \frac{\mathrm{pr}(y|x_1)(1 - \mathrm{pr}(y|x_1))}{n_{x_1}} \\
&\quad + \frac{1}{n_{x_2}} \left(\sum_u \mathrm{pr}(y|x_1, u)^2 \,\mathrm{pr}(u|x_2) - \left(\sum_u \mathrm{pr}(y|x_1, u)\, \mathrm{pr}(u|x_2) \right)^2 \right),
\end{aligned} \tag{5}$$

respectively. The derivations are provided in the Appendix.

Since these involve expectations of fractionals in the variances given by equations (4) and (5), it is difficult to derive closed-form approximations of their exact variances. To avoid this difficulty, Elandt-Johnson and Johnson [5] introduced several approximated expectations of fractionals based on the delta method (Anderson [1], Oehlert [14], Ver Hoef [31]). Intuitively, the delta method is based on Taylor's series expansion for the function of parameters and often provides a good approximation of variance estimates. Assuming that n_x is sufficiently large, we use one of their formulas as an approximation of our variances:

$$E\left(\frac{1}{n_{x,u}}\right) \simeq \frac{1}{n_x \,\mathrm{pr}(u|x)}.$$

Then, we have

$$E\left(\frac{n_{x_2,u}^2}{n_{x_1,u}}\right) \simeq \frac{n_{x_2}^2}{n_{x_1}\,\mathrm{pr}(u|x_1)}\left(\frac{\mathrm{pr}(u|x_2)(1-\mathrm{pr}(u|x_2))}{n_{x_2}} + \mathrm{pr}(u|x_2)^2\right).$$

4.3. Variable selection

In this section, when both X and Y are dichotomous variables, letting $\mathbf{U} = \mathbf{S} \cup \mathbf{Z}$ and $\mathbf{T} = \mathbf{W} \cup \mathbf{R}$ for discrete intermediate variables $\mathbf{S}$ and $\mathbf{W}$ and discrete covariates $\mathbf{Z}$ and $\mathbf{R}$, we assume that both $\mathrm{NDE}_y^S(x_1, x_2; \mathbf{Z})$ (and $\mathrm{NIE}_y^S(x_1, x_2; \mathbf{Z})$) and $\mathrm{NDE}_y^W(x_1, x_2; \mathbf{R})$ (and $\mathrm{NIE}_y^W(x_1, x_2; \mathbf{R})$) are estimated by using sets of covariates $\mathbf{Z}$ and $\mathbf{R}$ respectively, under the identification conditions presented in Section 2.2. Then, when two non-empty sets of the variables $\mathbf{T}$ and $\mathbf{U}$ satisfy both $X \perp\!\!\!\perp \mathbf{U}|\mathbf{T}$ and $Y \perp\!\!\!\perp \mathbf{T}|\{X\} \cup \mathbf{U}$, $\mathbf{T}$ and $\mathbf{U}$ are weakly equivalent to each other. Thus, the variable selection problem, that is, whether it is better to use both sets of variables than just one to obtain a point estimator with a smaller variance, can be addressed from the viewpoint of weak equivalence.

Regarding this problem, under the identification conditions presented in Section 2.2, the following results are obtained.

Theorem 4. (I) *When we have both* $\mathrm{NDE}_y^S(x_1, x_2; \mathbf{Z}) = \mathrm{NDE}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})$ *and* $\mathrm{NIE}_y^S(x_1, x_2; \mathbf{Z}) = \mathrm{NIE}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})$ *and the condition* $Y \perp\!\!\!\perp \mathbf{T}|\{X\} \cup \mathbf{U}$ *hold for the available data*, *we have*

$$\mathrm{a.var}\left\{\widehat{\mathrm{NDE}}_y^S(x_1, x_2; \mathbf{Z})\right\} \le \mathrm{a.var}\left\{\widehat{\mathrm{NDE}}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})\right\} \tag{6}$$

for the NDE, *and*

$$\mathrm{a.var}\left\{\widehat{\mathrm{NIE}}_y^S(x_1, x_2; \mathbf{Z})\right\} \le \mathrm{a.var}\left\{\widehat{\mathrm{NIE}}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})\right\} \tag{7}$$

for the NIE.

(II) *When we have both* $\mathrm{NDE}_y^W(x_1, x_2; \mathbf{R}) = \mathrm{NDE}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})$ *and* $\mathrm{NIE}_y^W(x_1, x_2; \mathbf{R}) = \mathrm{NIE}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})$ *and the condition* $X \perp\!\!\!\perp \mathbf{U}|\mathbf{T}$ *holds for the available data*, *we have*

$$\mathrm{a.var}\left\{\widehat{\mathrm{NDE}}_y^W(x_1, x_2; \mathbf{R})\right\} \le \mathrm{a.var}\left\{\widehat{\mathrm{NDE}}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})\right\} \tag{8}$$

for the NDE if both

$$1 + n_{x_2}\,\mathrm{pr}(t|x_2) \le n_{x_1}\,\mathrm{pr}(t|x_1)$$

and

$$\mathrm{cov}(t) = \sum_{s} \mathrm{pr}(y_1|x_1, t, u)\,\mathrm{pr}(y_1|x_2, t, u)\,\mathrm{pr}(u|t) - \mathrm{pr}(y_1|x_1, t)\,\mathrm{pr}(y_1|x_2, t) \le 0$$

hold for any t. In addition, we have

$$\mathrm{a.var}\bigl\{\widehat{\mathrm{NIE}}_y^{W}(x_1, x_2; \mathbf{R})\bigr\} \le \mathrm{a.var}\bigl\{\widehat{\mathrm{NIE}}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})\bigr\} \tag{9}$$

for the NIE if

$$1 + n_{x_2}\,\mathrm{pr}(t|x_2) \le n_{x_1}\,\mathrm{pr}(t|x_1)$$

hold for any t.

The proofs for (I) and (II) of Theorem 4 are provided in the Appendix. Compared with the results of linear regression models, Theorem 4(I) is as expected: control for additional intermediate variables and covariates that are directly associated with the treatment (not with the response directly) will increase the variance or leave it unchanged. The surprising result is Theorem 4(II): controlling for intermediate variables and covariates that are directly associated with the response (not with the treatment directly) may increase the variance when $1 + n_{x_2}\,\mathrm{pr}(t|x_2) \le n_{x_1}\,\mathrm{pr}(t|x_1)$ holds for any t. In some ways, this result shows a "negative" relationship in the sense that controlling for intermediate variables and covariates directly associated with the response may turn out to increase the variances of the NDE and NIE. This property is contrary to the case of linear regression models, because the variance of the regression coefficient is always decreasing (asymptotically) under such a situation (e.g., Clogg *et al.* [3], Kuroki and Cai [10], Kuroki and Miyakawa [13], Wermuth [33]).

5. Simulation experiments

We compare the variances described in Sections 4.2 and 4.3 through simulation experiments. For simplicity, we consider only the case where both $X \perp\!\!\!\perp S|W$ and $Y \perp\!\!\!\perp W|\{X, S\}$ hold, and there are two observed dichotomous intermediate variables S and W. This situation can be described by the directed acyclic graph wherein S_1 and S_2 in Figure 2 are replaced by W and S respectively, and S and W are weakly equivalent to each other from Theorem 1.

The setting of conditional probabilities of S given W and W given X are fixed at $\mathrm{pr}(s_1|w_1) = 0.7$, $\mathrm{pr}(s_1|w_2) = 0.2$, $\mathrm{pr}(w_1|x_1) = 0.8$ and $\mathrm{pr}(w_1|x_2) = 0.2$. In addition, letting:

(A.1) $\mathrm{pr}(y_1|x_1, s_1) = 0.7$, $\mathrm{pr}(y_1|x_1, s_2) = 0.2$, $\mathrm{pr}(y_1|x_2, s_1) = 0.6$, $\mathrm{pr}(y_1|x_2, s_2) = 0.2$;
(A.2) $\mathrm{pr}(y_1|x_1, s_1) = 0.7$, $\mathrm{pr}(y_1|x_1, s_2) = 0.2$, $\mathrm{pr}(y_1|x_2, s_1) = 0.2$, $\mathrm{pr}(y_1|x_2, s_2) = 0.6$;
(B.1) $\mathrm{pr}(x_1) = 0.1$; (B.2) $\mathrm{pr}(x_1) = 0.5$; (B.3) $\mathrm{pr}(x_1) = 0.9$,

we consider the following six scenarios in accordance with the description given in Section 4.2:

1. Setting (A.1) + (B.1): the case where both $n_{x_2}\,\mathrm{pr}(w|x_2)+1 \geq n_{x_1}\,\mathrm{pr}(w|x_1)$ and $\mathrm{cov}(w) \geq 0$ hold true for any $w \in \{w_1, w_2\}$.
2. Setting (A.1) + (B.2): the case where $\mathrm{pr}(x_1)=\mathrm{pr}(x_2)=0.5$ and $\mathrm{cov}(w) \geq 0$ hold true for any $w \in \{w_1, w_2\}$.
3. Setting (A.1) + (B.3): the case where both $n_{x_2}\,\mathrm{pr}(w|x_2)+1 \leq n_{x_1}\,\mathrm{pr}(w|x_1)$ and $\mathrm{cov}(w) \geq 0$ hold true for any $w \in \{w_1, w_2\}$.
4. Setting (A.2) + (B.1): the case where both $n_{x_2}\,\mathrm{pr}(w|x_2)+1 \geq n_{x_1}\,\mathrm{pr}(w|x_1)$ and $\mathrm{cov}(w) \leq 0$ hold true for any $w \in \{w_1, w_2\}$.
5. Setting (A.2) + (B.2): the case where $\mathrm{pr}(x_1)=\mathrm{pr}(x_2)=0.5$ and $\mathrm{cov}(w) \leq 0$ hold true for any $w \in \{w_1, w_2\}$.
6. Setting (A.2) + (B.3): the case where both $n_{x_2}\,\mathrm{pr}(w|x_2)+1 \leq n_{x_1}\,\mathrm{pr}(w|x_1)$ and $\mathrm{cov}(w) \leq 0$ hold true for any $w \in \{w_1, w_2\}$.

Table 4 represents the variance estimates from 10 000 replications for sample size $N = 1000$ and 2000. Columns labeled "S" show the variances when an intermediate variable S is used to estimate the NDE and NIE, columns labeled "W" show the variances when an intermediate variable W is used to estimate the NDE and NIE, and columns labeled "$\{S, W\}$" show the variances

Table 4. Simulation results comparing the variances with the asymptotic variance

			(A.1) + (B.1)			(A.1) + (B.2)			(A.1) + (B.3)		
			S	W	$\{S, W\}$	S	W	$\{S, W\}$	S	W	$\{S, W\}$
NDE	$n = 1000$	$\sqrt{\text{a.var}}$	0.0498	0.0842	0.0759	0.0288	0.0423	0.0386	0.0460	0.0537	0.0510
		$\sqrt{\text{var}}$	0.0506	0.0864	0.0810	0.0288	0.0423	0.0386	0.0458	0.0534	0.0502
	$n = 2000$	$\sqrt{\text{a.var}}$	0.0352	0.0595	0.0536	0.0203	0.0299	0.0272	0.0325	0.0379	0.0356
		$\sqrt{\text{var}}$	0.0351	0.0597	0.0548	0.0204	0.0301	0.0272	0.0329	0.0383	0.0356
NIE	$n = 1000$	$\sqrt{\text{a.var}}$	0.0365	0.0708	0.0679	0.0190	0.0319	0.0319	0.0259	0.0256	0.0340
		$\sqrt{\text{var}}$	0.0375	0.0737	0.0727	0.0191	0.0322	0.0322	0.0260	0.0256	0.0325
	$n = 2000$	$\sqrt{\text{a.var}}$	0.0258	0.0500	0.0480	0.0134	0.0226	0.0225	0.0183	0.0181	0.0234
		$\sqrt{\text{var}}$	0.0258	0.0503	0.0491	0.0134	0.0227	0.0225	0.0183	0.0181	0.0229
			(A.2) + (B.1)			(A.2) + (B.2)			(A.2) + (B.3)		
			S	W	$\{S, W\}$	S	W	$\{S, W\}$	S	W	$\{S, W\}$
NDE	$n = 1000$	$\sqrt{\text{a.var}}$	0.0520	0.0846	0.0773	0.0350	0.0438	0.0434	0.0642	0.0593	0.0678
		$\sqrt{\text{var}}$	0.0523	0.0863	0.0830	0.0352	0.0439	0.0436	0.0639	0.0592	0.0682
	$n = 2000$	$\sqrt{\text{a.var}}$	0.0368	0.0598	0.0546	0.0248	0.0310	0.0307	0.0454	0.0419	0.0476
		$\sqrt{\text{var}}$	0.0372	0.0616	0.0567	0.0246	0.0310	0.0306	0.0454	0.0419	0.0473
NIE	$n = 1000$	$\sqrt{\text{a.var}}$	0.0365	0.0708	0.0679	0.0190	0.0319	0.0319	0.0259	0.0256	0.0340
		$\sqrt{\text{var}}$	0.0372	0.0734	0.0738	0.0190	0.0318	0.0319	0.0260	0.0256	0.0330
	$n = 2000$	$\sqrt{\text{a.var}}$	0.0258	0.0500	0.0480	0.0134	0.0226	0.0225	0.0183	0.0181	0.0234
		$\sqrt{\text{var}}$	0.0264	0.0516	0.0502	0.0134	0.0226	0.0225	0.0183	0.0181	0.0228

when both S and W are used to estimate the NDE and the NIE. The first rows show the square root value of the asymptotic variance calculated from the equations in Section 4.2, denoted as $\sqrt{\text{a.var}}$, and the second rows show the square root value of the variance obtained from simulation experiments, denoted as $\sqrt{\text{var}}$. From Table 4, we draw the following conclusions.

1. The ratio of the variance to the asymptotic variance is between 0.92 and 1.05 for all settings, which indicates that the asymptotic variances seem to be reasonable approximations.
2. In settings for both the NDE and NIE, the variance when S is selected is smaller than the variance when $\{S, W\}$ is selected, which is consistent with Theorem 4(I).
3. In the case of the NIE, the variance when W is selected is smaller than the variance when $\{S, W\}$ is selected for all settings involving (B.3), which is consistent with Theorem 4(II). This indicates that it is not always better to use all the available variable information to estimate the NIE. In addition, the variance when W is selected is smaller than the variance when S is selected, which indicates that selecting a set of variables that has a direct effect on a response cannot always improve the estimation accuracy of the NIE.
4. For settings involving (A.2) of the NDE, the order of the magnitude of the variances vary according to the intermediate variables used. Especially, in setting (A.2) + (B.3), for the NDE, the variances when W is selected are smaller than the variances when $\{S, W\}$ is selected, which is theoretically predictable from Theorem 4(II).
5. The performances of the NIE for setting (A.1) are almost the same as those for setting (A.2), because the information on $\text{pr}(y_1|x_2, s)(s \in \{s_1, s_2\})$ is not used to estimate the NIE in the simulation experiments.

6. Discussion

6.1. Conclusion

This paper introduced the new concept of weak equivalence wherein two different sets of variables estimate the same direct and indirect effect, and the sufficient conditions for weak equivalence between two sets of variables were provided. The concept of equivalence can help us choose intermediate variables and covariates, and thus reduce costs without amplifying the bias related to the target quantities. In addition, we discussed the variable selection problem from the viewpoint of estimation accuracy when two sets of variables are weakly equivalent to each other. Finally, through simulation experiments, we demonstrated the paradox that selecting a set of variables that has a direct effect on a response cannot always improve the estimation accuracy, which is a similar phenomenon described by Kuroki and Cai [11] and Robinson and Jewell [22], but contrary to the situation found in linear regression models (e.g., Kuroki and Cai [10], Kuroki and Miyakawa [13]). In this paper, we transformed the NDE and NIE to standardized quantities based on the subpopulation $X = x'$ using the exchangeability between marginal probabilities and conditional probabilities from the assumption of randomized experiments for the treatment X. It would be possible to derive the variance estimators of the NDE and NIE without using the condition of exchangeability; however, the derivation has been omitted due to its complexity. Nevertheless, our results are still valuable in the sense that this paper draws attention to the fact that the observation in linear regression analysis does not always hold for other statistical measures.

6.2. Future work

In this section, we would like to point out some future work. First, although we focused on sufficient conditions for weak equivalence, it would be possible to derive necessary and sufficient conditions through precise parameter tuning. The derivation of the conditions would be important from mathematical viewpoint and would be useful in the sense that it makes clear that there are situations where two sets of variables are weakly equivalent to each other but the proposed sufficient conditions do not hold. However, we would like to leave the discussion of whether necessary and sufficient conditions through such parameter tuning are practical or not as future work. Second, it is noted that the NDE or NIE used in this paper are not variation independent of margin such as functions of the odds ratio (Edwards [4], Wermuth *et al.* [34]). Thus, the discussion based on the function of odds ratio would also be future work.

Appendix

Equations (4) and (5)

Letting $U = \mathbf{S} \cup \mathbf{Z}$, based on the variance basic formula, we formulate the variance of the NDE as

$$
\begin{aligned}
&\mathrm{var}\{\widehat{\mathrm{NDE}}_y^S(x_1, x_2; \mathbf{Z})\} \\
&= \mathrm{var}\{E(\widehat{\mathrm{NDE}}_y^S(x_1, x_2; \mathbf{Z})|n_{x_1,u}, n_{x_2,u})\} + E\{\mathrm{var}(\widehat{\mathrm{NDE}}_y^S(x_1, x_2; \mathbf{Z})|n_{x_1,u}, n_{x_2,u})\} \\
&= \mathrm{var}\left\{\sum_u (\mathrm{pr}(y|x_1, u) - \mathrm{pr}(y|x_2, u))\widehat{\mathrm{pr}}(u|x_2)\right\} \\
&\quad + \sum_u E\left\{\frac{\mathrm{pr}(y|x_1, u)(1 - \mathrm{pr}(y|x_1, u))}{n_{x_1,u}}\widehat{\mathrm{pr}}(u|x_2)^2 + \frac{\mathrm{pr}(y|x_2, u)(1 - \mathrm{pr}(y|x_2, u))}{n_{x_2,u}}\widehat{\mathrm{pr}}(u|x_2)^2\right\} \\
&= \sum_u \{\mathrm{pr}(y|x_1, u) - \mathrm{pr}(y|x_2, u)\}^2 \mathrm{var}\{\widehat{\mathrm{pr}}(u|x_2)\} \\
&\quad + \sum_{u \neq u'} \{\mathrm{pr}(y|x_1, u) - \mathrm{pr}(y|x_2, u)\}\{\mathrm{pr}(y|x_1, u') - \mathrm{pr}(y|x_2, u')\} \mathrm{cov}\{\widehat{\mathrm{pr}}(u|x_2), \widehat{\mathrm{pr}}(u'|x_2)\} \\
&\quad + \sum_u \frac{\mathrm{pr}(y|x_1, u)(1 - \mathrm{pr}(y|x_1, u))}{n_{x_2}^2} E\left(\frac{n_{x_2,u}^2}{n_{x_1,u}}\right) + \sum_u \frac{\mathrm{pr}(y|x_2, u)(1 - \mathrm{pr}(y|x_2, u))}{n_{x_2}^2} E(n_{x_2,u}) \\
&= \sum_u \{\mathrm{pr}(y|x_1, u) - \mathrm{pr}(y|x_2, u)\}^2 \frac{\mathrm{pr}(u|x_2)(1 - \mathrm{pr}(u|x_2))}{n_{x_2}} \\
&\quad - \sum_{u \neq u'} \{\mathrm{pr}(y|x_1, u) - \mathrm{pr}(y|x_2, u)\}\{\mathrm{pr}(y|x_1, u') - \mathrm{pr}(y|x_2, u')\} \frac{\mathrm{pr}(u|x_2)\,\mathrm{pr}(u'|x_2)}{n_{x_2}}
\end{aligned}
$$

$$+\sum_u \frac{\mathrm{pr}(y|x_1,u)(1-\mathrm{pr}(y|x_1,u))}{n_{x_2}^2} E\left(\frac{n_{x_2,u}^2}{n_{x_1,u}}\right) + \sum_u \frac{\mathrm{pr}(y|x_2,u)(1-\mathrm{pr}(y|x_2,u))}{n_{x_2}} \mathrm{pr}(u|x_2)$$

$$= \sum_u \{\mathrm{pr}(y|x_1,u) - \mathrm{pr}(y|x_2,u)\}^2 \frac{\mathrm{pr}(u|x_2)}{n_{x_2}} - \frac{\mathrm{NDE}_y^{S2}(x_1,x_2;\mathbf{Z})}{n_{x_2}}$$

$$+\sum_u \frac{\mathrm{pr}(y|x_1,u)(1-\mathrm{pr}(y|x_1,u))}{n_{x_2}^2} E\left(\frac{n_{x_2,u}^2}{n_{x_1,u}}\right) + \sum_u \frac{\mathrm{pr}(y|x_2,u)(1-\mathrm{pr}(y|x_2,u))}{n_{x_2}} \mathrm{pr}(u|x_2).$$

Thus, we obtain equation (4).

Similarly, the variance of the NIE is formulated as

$$\mathrm{var}\{\widehat{\mathrm{NIE}}_y^S(x_1,x_2;\mathbf{Z})\}$$

$$= \mathrm{var}\{E(\widehat{\mathrm{NIE}}_y^S(x_1,x_2;\mathbf{Z})|n_{x_1,u},n_{x_2,u})\} + E\{\mathrm{var}(\widehat{\mathrm{NIE}}_y^S(x_1,x_2;\mathbf{Z})|n_{x_1,u},n_{x_2,u})\}$$

$$= \mathrm{var}\left\{\sum_u \mathrm{pr}(y|x_1,u)\big(\widehat{\mathrm{pr}}(u|x_1) - \widehat{\mathrm{pr}}(u|x_2)\big)\right\}$$

$$+ E\left\{\sum_u \frac{\mathrm{pr}(y|x_1,u)(1-\mathrm{pr}(y|x_1,u))}{n_{x_1,u}} \big(\widehat{\mathrm{pr}}(u|x_1) - \widehat{\mathrm{pr}}(u|x_2)\big)^2\right\}$$

$$= \sum_u \mathrm{pr}(y|x_1,u)\big(1-\mathrm{pr}(y|x_1,u)\big) E\left\{\frac{\widehat{\mathrm{pr}}(u|x_2)^2 - 2\widehat{\mathrm{pr}}(u|x_2)\widehat{\mathrm{pr}}(u|x_1) + \widehat{\mathrm{pr}}(u|x_1)^2}{n_{x_1,u}}\right\}$$

$$+ \sum_u \mathrm{pr}(y|x_1,u)^2\, \mathrm{var}\{\widehat{\mathrm{pr}}(u|x_2) - \widehat{\mathrm{pr}}(u|x_1)\}$$

$$+ \sum_{u\neq u'} \mathrm{pr}(y|x_1,u)\,\mathrm{pr}(y|x_1,u')\,\mathrm{cov}\big(\widehat{\mathrm{pr}}(u|x_1) - \widehat{\mathrm{pr}}(u|x_2), \widehat{\mathrm{pr}}(u'|x_1) - \widehat{\mathrm{pr}}(u'|x_2)\big)$$

$$= \sum_u \mathrm{pr}(y|x_1,u)\big(1-\mathrm{pr}(y|x_1,u)\big)\left\{E\left\{\frac{\widehat{\mathrm{pr}}(u|x_2)^2}{n_{x_1,u}}\right\} - 2\frac{E(\widehat{\mathrm{pr}}(u|x_2))}{n_{x_1}} + \frac{E(\widehat{\mathrm{pr}}(u|x_1))}{n_{x_1}}\right\}$$

$$+ \sum_u \mathrm{pr}(y|x_1,u)^2 \left\{\frac{\mathrm{pr}(u|x_1)(1-\mathrm{pr}(u|x_1))}{n_{x_1}} + \frac{\mathrm{pr}(u|x_2)(1-\mathrm{pr}(u|x_2))}{n_{x_2}}\right\}$$

$$- \sum_{u\neq u'} \mathrm{pr}(y|x_1,u)\mathrm{pr}(y|x_1,u')\left(\frac{\mathrm{pr}(u|x_1)\,\mathrm{pr}(u'|x_1)}{n_{x_1}} + \frac{\mathrm{pr}(u|x_2)\,\mathrm{pr}(u'|x_2)}{n_{x_2}}\right)$$

$$= \sum_u \mathrm{pr}(y|x_1,u)\big(1-\mathrm{pr}(y|x_1,u)\big)\left\{E\left(\frac{\widehat{\mathrm{pr}}(u|x_2)^2}{n_{x_1,u}}\right) - 2\frac{\mathrm{pr}(u|x_2)}{n_{x_1}} + \frac{\mathrm{pr}(u|x_1)}{n_{x_1}}\right\}$$

$$+ \sum_u \frac{\mathrm{pr}(y|x_1,u)^2\,\mathrm{pr}(u|x_1)}{n_{x_1}} + \sum_u \frac{\mathrm{pr}(y|x_1,u)^2\,\mathrm{pr}(u|x_2)}{n_{x_2}} - \frac{(\sum_u \mathrm{pr}(y|x_1,u)\,\mathrm{pr}(u|x_1))^2}{n_{x_1}}$$

$$
\begin{aligned}
&- \frac{(\sum_u \mathrm{pr}(y|x_1,u)\,\mathrm{pr}(u|x_2))^2}{n_{x_2}} \\
&= \frac{\mathrm{pr}(y|x_1)(1-\mathrm{pr}(y|x_1))}{n_{x_1}} + \sum_u \frac{\mathrm{pr}(y|x_1,u)(1-\mathrm{pr}(y|x_1,u))}{n_{x_2}^2} E\left(\frac{n_{x_2,u}^2}{n_{x_1,u}}\right) \\
&\quad + \frac{1}{n_{x_2}}\left(\sum_u \mathrm{pr}(y|x_1,u)^2\,\mathrm{pr}(u|x_2) - \left(\sum_u \mathrm{pr}(y|x_1,u)\,\mathrm{pr}(u|x_2)\right)^2\right) \\
&\quad - 2\sum_u \mathrm{pr}(y|x_1,u)\big(1-\mathrm{pr}(y|x_1,u)\big)\frac{\mathrm{pr}(u|x_2)}{n_{x_1}}.
\end{aligned}
$$

Thus, we obtain equation (5).

Equations (6) and (7)

In this section, letting $U = \mathbf{S} \cup \mathbf{Z}$ and $T = \mathbf{R} \cup \mathbf{W}$, we compare the variance of $\widehat{\mathrm{NDE}}_y^S(x_1, x_2; \mathbf{Z})$ with that of $\widehat{\mathrm{NDE}}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})$ under the condition $Y \perp\!\!\!\perp T|\{X, U\}$. Noting that $\mathrm{pr}(y|x_i, u, t) = \mathrm{pr}(y|x_i, u)$ $(i = 1, 2)$, we have

$$
\begin{aligned}
&\mathrm{a.var}\big\{\widehat{\mathrm{NDE}}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})\big\} - \mathrm{a.var}\big\{\widehat{\mathrm{NDE}}_y^{S}(x_1, x_2; \mathbf{Z})\big\} \\
&\quad = \sum_u \mathrm{pr}(y|x_1,u)\big(1-\mathrm{pr}(y|x_1,u)\big)\left\{\sum_t E\left(\frac{n_{x_2,u,t}^2}{n_{x_1,u,t}}\right) - E\left(\frac{n_{x_2,u}^2}{n_{x_1,u}}\right)\right\} \\
&\quad = \sum_u \mathrm{pr}(y|x_1,u)\big(1-\mathrm{pr}(y|x_1,u)\big) E\left(\sum_t \frac{n_{x_2,u,t}^2}{n_{x_1,u,t}} - \frac{n_{x_2,u}^2}{n_{x_1,u}}\right).
\end{aligned}
$$

From the Cauchy–Schwarz inequality, for $n_{x_1,4,t} \neq 0$, since we obtain

$$
n_{x_1,u}\sum_t \frac{n_{x_2,u,t}^2}{n_{x_1,u,t}} = \sum_t n_{x_1,u,t} \sum_t \frac{n_{x_2,u,t}^2}{n_{x_1,u,t}} \geq \left(\sum_t n_{x_2,u,t}\right)^2 = n_{x_2,u}^2,
$$

we obtain

$$
E\left(\sum_t \frac{n_{x_2,u,t}^2}{n_{x_1,u,t}} - \frac{n_{x_2,u}^2}{n_{x_1,u}}\right) \geq 0. \tag{10}
$$

Thus, $\mathrm{a.var}\{\widehat{\mathrm{NDE}}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})\} \geq \mathrm{a.var}\{\widehat{\mathrm{NDE}}_y^{S}(x_1, x_2; \mathbf{Z})\}$ holds.

By the similar procedure, $\mathrm{a.var}\{\widehat{\mathrm{NIE}}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})\} \geq \mathrm{a.var}\{\widehat{\mathrm{NIE}}_y^{S}(x_1, x_2; \mathbf{Z})\}$ can be obtained.

Equations (8) and (9)

First, $U = \mathbf{S} \cup \mathbf{Z}$ and $T = \mathbf{R} \cup \mathbf{W}$, we compare the variance of $\widehat{\mathrm{NDE}}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})$ with that of $\widehat{\mathrm{NDE}}_y^{W}(x_1, x_2; \mathbf{R})$ under the condition $U \perp\!\!\!\perp X|T$. Then, we have

$$
\begin{aligned}
&\mathrm{a.var}\{\widehat{\mathrm{NDE}}_y^{S,W}(x_1, x_2; \mathbf{Z}, \mathbf{R})\} - \mathrm{a.var}\{\widehat{\mathrm{NDE}}_y^{W}(x_1, x_2; \mathbf{R})\} \\
&= \sum_t \left\{ \sum_u \big(\mathrm{pr}(y|x_1, u, t) - \mathrm{pr}(y|x_2, u, t)\big)^2 \frac{\mathrm{pr}(u, t|x_2)}{n_{x_2}} \big(\mathrm{pr}(y|x_1, t) - \mathrm{pr}(y|x_2, t)\big)^2 \frac{\mathrm{pr}(t|x_2)}{n_{x_2}} \right\} \\
&\quad + \sum_t \left\{ \sum_u \frac{\mathrm{pr}(y|x_2, u, t)(1 - \mathrm{pr}(y|x_2, u, t))}{n_{x_2}} \mathrm{pr}(u, t|x_2) \right. \\
&\qquad \left. - \frac{\mathrm{pr}(y|x_2, t)(1 - \mathrm{pr}(y|x_2, t))}{n_{x_2}} \mathrm{pr}(t|x_2) \right\} \\
&\quad + \sum_t \left\{ \sum_u \frac{\mathrm{pr}(y|x_1, u, t)(1 - \mathrm{pr}(y|x_1, u, t))}{n_{x_2}^2} E\left(\frac{n_{x_2,u,t}^2}{n_{x_1,u,t}}\right) \right. \\
&\qquad \left. - \frac{\mathrm{pr}(y|x_1, t)(1 - \mathrm{pr}(y|x_1, t))}{n_{x_2}^2} E\left(\frac{n_{x_2,t}^2}{n_{x_1,t}}\right) \right\} \\
&= \sum_t \left\{ \sum_u \big(\mathrm{pr}(y|x_1, u, t) - \mathrm{pr}(y|x_2, u, t)\big)^2 \mathrm{pr}(u|t) - \big(\mathrm{pr}(y|x_1, t) - \mathrm{pr}(y|x_2, t)\big)^2 \right\} \frac{\mathrm{pr}(t|x_2)}{n_{x_2}} \\
&\quad + \sum_t \left\{ \sum_u \mathrm{pr}(y|x_2, u, t)\big(1 - \mathrm{pr}(y|x_2, u, t)\big) \mathrm{pr}(u|t) \right. \\
&\qquad \left. - \mathrm{pr}(y|x_2, t)\big(1 - \mathrm{pr}(y|x_2, t)\big) \right\} \frac{\mathrm{pr}(t|x_2)}{n_{x_2}} \\
&\quad + \sum_t \left\{ \sum_u \frac{\mathrm{pr}(y|x_1, u, t)(1 - \mathrm{pr}(y|x_1, u, t))}{n_{x_2}^2} E\left(\frac{n_{x_2,u,t}^2}{n_{x_1,u,t}}\right) \right. \\
&\qquad \left. - \frac{\mathrm{pr}(y|x_1, t)(1 - \mathrm{pr}(y|x_1, t))}{n_{x_2}^2} E\left(\frac{n_{x_2,t}^2}{n_{x_1,t}}\right) \right\} \\
&= \sum_t \left\{ \sum_u \mathrm{pr}(y|x_1, u, t)^2 \mathrm{pr}(u|t) - \mathrm{pr}(y|x_1, t)^2 \right. \\
&\qquad \left. + 2\left\{ \mathrm{pr}(y|x_1, t)\,\mathrm{pr}(y|x_2, t) - \sum_u \mathrm{pr}(y|x_1, u, t)\,\mathrm{pr}(y|x_2, u, t)\,\mathrm{pr}(u|t) \right\} \right\} \frac{\mathrm{pr}(t|x_2)}{n_{x_2}} \\
&\qquad + \sum_t \left\{ \sum_u \frac{\mathrm{pr}(y|x_1, u, t)(1 - \mathrm{pr}(y|x_1, u, t))}{n_{x_2}^2} E\left(\frac{n_{x_2,u,t}^2}{n_{x_1,u,t}}\right) \right.
\end{aligned}
$$

$$
-\frac{\text{pr}(y|x_1,t)(1-\text{pr}(y|x_1,t))}{n_{x_2}^2}E\left(\frac{n_{x_2,t}^2}{n_{x_1,t}}\right)\Bigg\}
$$

$$
\begin{aligned}
&=\sum_t\Bigg\{\sum_u\frac{\text{pr}(y|x_1,u,t)(1-\text{pr}(y|x_1,u,t))}{n_{x_2}^2}\left(E\left(\frac{n_{x_2,u,t}^2}{n_{x_1,u,t}}\right)-\text{pr}(u|t)E\left(\frac{n_{x_2,t}^2}{n_{x_1,t}}\right)\right)\\
&\quad+\left(\sum_u\text{pr}(y|x_1,u,t)\big(1-\text{pr}(y|x_1,u,t)\big)\,\text{pr}(u|t)-\text{pr}(y|x_1,t)\big(1-\text{pr}(y|x_1,t)\big)\right)\\
&\quad\times\left(\frac{1}{n_{x_2}^2}E\left(\frac{n_{x_2,t}^2}{n_{x_1,t}}\right)-\frac{\text{pr}(t|x_2)}{n_{x_2}}\right)\\
&\quad+2\left\{\text{pr}(y|x_1,t)\,\text{pr}(y|x_2,t)-\sum_u\text{pr}(y|x_1,u,t)\,\text{pr}(y|x_2,u,t)\,\text{pr}(u|t)\right\}\frac{\text{pr}(t|x_2)}{n_{x_2}}\Bigg\}.
\end{aligned}
$$

Here, since $\text{cov}(t)=\sum_u\text{pr}(y|x_1,u,t)\,\text{pr}(y|x_2,u,t)\,\text{pr}(u|t)-\text{pr}(y|x_1,t)\,\text{pr}(y|x_2,t)\leq 0$ under the conditions, the third term is non-negative. In addition, noting that we approximate

$$
E\left(\frac{1}{n_{x_1,t}}\right)\simeq\frac{1}{n_{x_1}\,\text{pr}(t|x_1)},\qquad E\left(\frac{1}{n_{x_1,u,t}}\right)\simeq\frac{1}{n_{x_1}\,\text{pr}(u,t|x_1)},
$$

from the assumption $1+(n_{x_2}-1)\,\text{pr}(t|x_2)\leq 1+n_{x_2}\,\text{pr}(t|x_2)\leq n_{x_1}\,\text{pr}(t|x_1)$, we have

$$
\frac{1}{n_{x_2}}E\left(\frac{n_{x_2,t}^2}{n_{x_1,t}}\right)-\text{pr}(t|x_2)\simeq\frac{\text{pr}(t|x_2)}{n_{x_1}\,\text{pr}(t|x_1)}\big(1+(n_{x_2}-1)\,\text{pr}(t|x_2)-n_{x_1}\,\text{pr}(t|x_1)\big)\leq 0. \tag{11}
$$

Thus, since the second term is non-negative, we have $\text{a.var}\{\widehat{\text{NDE}}_y^{S,W}(x_1,x_2;\mathbf{Z},\mathbf{R})\}\geq\text{a.var}\{\widehat{\text{NDE}}_y^{W}(x_1,x_2;\mathbf{R})\}$.

Next, we compare the variance of $\widehat{\text{NIE}}_y^{S,W}(x_1,x_2;\mathbf{Z},\mathbf{R})$ with that of $\widehat{\text{NIE}}_y^{W}(x_1,x_2;\mathbf{R})$ under the condition $U \perp\!\!\!\perp X|T$.

From equations (10) and (11), we have

$$
\begin{aligned}
&\text{var}\big\{\widehat{\text{NIE}}_y^{S,W}(x_1,x_2;\mathbf{Z},\mathbf{R})\big\}-\text{var}\big\{\widehat{\text{NIE}}_y^{W}(x_1,x_2;\mathbf{R})\big\}\\
&\quad=\frac{2}{n_{x_1}}\sum_t\Bigg(\text{pr}(y|x_1,t)\big(1-\text{pr}(y|x_1,t)\big)\\
&\qquad\qquad-\sum_u\text{pr}(y|x_1,u,t)\big(1-\text{pr}(y|x_1,u,t)\big)\,\text{pr}(u|t)\Bigg)\,\text{pr}(t|x_2)\\
&\qquad+\sum_t\Bigg(\sum_u\frac{\text{pr}(y|x_1,u,t)(1-\text{pr}(y|x_1,u,t))}{n_{x_2}^2}E\left(\frac{n_{x_2,u,t}^2}{n_{x_1,u,t}}\right)\\
&\qquad\qquad-\frac{\text{pr}(y|x_1,t)(1-\text{pr}(y|x_1,t))}{n_{x_2}^2}E\left(\frac{n_{x_2,t}^2}{n_{x_1,t}}\right)\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{1}{n_{x_2}} \sum_t \left(\sum_u \mathrm{pr}(y|x_1,u,t)^2 \,\mathrm{pr}(u|t) - \mathrm{pr}(y|x_1,t)^2 \right) \mathrm{pr}(t|x_2) \\
\geq{}& \frac{2}{n_{x_1}} \sum_t \Bigl(\mathrm{pr}(y|x_1,t)\bigl(1-\mathrm{pr}(y|x_1,t)\bigr) \\
&\qquad - \sum_u \mathrm{pr}(y|x_1,u,t)\bigl(1-\mathrm{pr}(y|x_1,u,t)\bigr)\,\mathrm{pr}(u|t) \Bigr) \mathrm{pr}(t|x_2) \\
&+ \sum_t \Biggl(\sum_u \frac{\mathrm{pr}(y|x_1,u,t)(1-\mathrm{pr}(y|x_1,u,t))}{n_{x_2}^2} \mathrm{pr}(u|t) E\left(\frac{n_{x_2,t}^2}{n_{x_1,t}} \right) \\
&\qquad - \frac{\mathrm{pr}(y|x_1,t)(1-\mathrm{pr}(y|x_1,t))}{n_{x_2}^2} E\left(\frac{n_{x_2,t}^2}{n_{x_1,t}} \right) \Biggr) \\
&+ \frac{1}{n_{x_2}} \sum_t \left(\sum_u \mathrm{pr}(y|x_1,u,t)^2 \,\mathrm{pr}(u|t) - \mathrm{pr}(y|x_1,t)^2 \right) \mathrm{pr}(t|x_2) \\
={}& \sum_t \left(\mathrm{pr}(y|x_1,t)\bigl(1-\mathrm{pr}(y|x_1,t)\bigr) - \sum_u \mathrm{pr}(y|x_1,u,t)\bigl(1-\mathrm{pr}(y|x_1,u,t)\bigr)\,\mathrm{pr}(u|t) \right) \\
&\times \left(\left(\frac{2}{n_{x_1}} + \frac{1}{n_{x_2}} \right) \mathrm{pr}(t|x_2) - \frac{1}{n_{x_2}^2} E\left(\frac{n_{x_2,t}^2}{n_{x_1,t}} \right) \right) \geq 0
\end{aligned}
$$

from equation (11). Thus, we obtain $\mathrm{var}\{\widehat{\mathrm{NIE}}_y^{S,W}(x_1,x_2;\mathbf{Z},\mathbf{R})\} \geq \mathrm{var}\{\widehat{\mathrm{NIE}}_y^{W}(x_1,x_2;\mathbf{R})\}$.

Acknowledgements

I thank two anonymous referees whose comments significantly improved the presentation of the paper. This paper was partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

[1] Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. *Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. MR1990662

[2] Cai, Z., Kuroki, M., Pearl, J. and Tian, J. (2008). Bounds on direct effects in the presence of confounded intermediate variables. *Biometrics* **64** 695–701. MR2526618

[3] Clogg, C.C., Petkova, E. and Shihadeh, E.S. (1992). Statistical methods for analyzing collapsibility in regression models. *J. Educ. Behav. Stat.* **17** 51–74.

[4] Edwards, A.W.F. (1963). The measure of association in a 2 × 2 table. *J. Roy. Statist. Soc. Ser. A* **126** 109–114.

[5] Elandt-Johnson, R.C. and Johnson, N.L. (1980). *Survival Models and Data Analysis*. New York: Wiley. MR0586940

[6] Hafeman, D.M. and Schwartz, S. (2009). Opening the black box: A motivation for the assessment of mediation. *Int*. *J*. *Epidemiol*. **38** 838–845.

[7] Imai, K., Keele, L. and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statist*. *Sci*. **25** 51–71. MR2741814

[8] Joffe, M.M. and Greene, T. (2009). Related causal frameworks for surrogate outcomes. *Biometrics* **65** 530–538. MR2751477

[9] Kaufman, S., Kaufman, J.S., MacLehose, R.F., Greenland, S. and Poole, C. (2005). Improved estimation of controlled direct effects in the presence of unmeasured confounding of intermediate variables. *Stat*. *Med*. **24** 1683–1702. MR2137644

[10] Kuroki, M. and Cai, Z. (2004). Selection of identifiability criteria for total effects by using path diagrams. *Uncertainty Artif*. *Intell*. **20** 333–340.

[11] Kuroki, M. and Cai, Z. (2011). Statistical analysis of 'probabilities of causation' using co-variate information. *Scand*. *J*. *Stat*. **38** 564–577. MR2833847

[12] Kuroki, M. and Miyakawa, M. (1999). Identifiability criteria for causal effects of joint interventions. *J*. *Japan Statist*. *Soc*. **29** 105–117. MR1765187

[13] Kuroki, M. and Miyakawa, M. (2003). Covariate selection for estimating the causal effect of control plans by using causal diagrams. *J*. *R*. *Stat*. *Soc*. *Ser*. *B Stat*. *Methodol*. **65** 209–222. MR1959822

[14] Oehlert, G.W. (1992). A note on the delta method. *Amer*. *Statist*. **46** 27–29. MR1149146

[15] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*: *Networks of Plausible Inference*. *The Morgan Kaufmann Series in Representation and Reasoning*. San Mateo, CA: Morgan Kaufmann. MR0965765

[16] Pearl, J. (2001). Direct and indirect effects. *Uncertainty Artif*. *Intell*. **17** 411–420.

[17] Pearl, J. (2009). *Causality*: *Models*, *Reasoning*, *and Inference*, 2nd ed. Cambridge: Cambridge Univ. Press. MR2548166

[18] Pearl, J. (2010). Confounding equivalence in causal inference. *Uncertainty Artif*. *Intell*. **26** 433–441.

[19] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Math*. *Modelling* **7** 1393–1512. MR0877758

[20] Robins, J.M. (1989). The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology*: *A Focus on AIDS*. (L. Sechrest, H. Freeman and A. Mulley, eds.) 113–159. Washington, DC: US Public Health Service, National Center for Health Services Research.

[21] Robins, J.M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3** 143–155.

[22] Robinson, L.D. and Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression models. *Int*. *Stat*. *Rev*. **59** 227–240.

[23] Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

[24] Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *J*. *Educ*. *Psychol*. **66** 688–701.

[25] Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann*. *Statist*. **6** 34–58. MR0472152

[26] Rubin, D.B. (1986). Which ifs have causal answers; comment on Holland (1986). *J*. *Amer*. *Statist*. *Assoc*. **81** 961–962.

[27] Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the* 21*st National Conference on Artificial Intelligence* 1219–1226.

[28] Technometrics Research Group (1999). The practice of graphical modelling. The Institute of Japanese Union of Scientists and Engineers.
[29] VanderWeele, T.J. (2011). Controlled direct and mediated effects: Definition, identification and bounds. *Scand*. *J*. *Stat*. **38** 551–563. MR2833846
[30] van der Laan, M.J. and Petersen, M.L. (2008). Direct effect models. *Int*. *J*. *Biostat*. **4** Art. 23. MR2456975
[31] Ver Hoef, J.M. (2012). Who invented the delta method? *Amer*. *Statist*. **66** 124–127. MR2968009
[32] Wang, Y. and Taylor, J.M.G. (2002). A measure of the proportion of treatment effect explained by a surrogate marker. *Biometrics* **58** 803–812. MR1945017
[33] Wermuth, N. (1989). Moderating effects in multivariate normal distributions. *Methodika* **3** 74–93.
[34] Wermuth, N., Marchetti, G.M. and Zwiernik, P. (2014). Binary distributions of concentric rings. *J*. *Multivariate Anal*. **130** 252–260. MR3229536

Received December 2013 and revised May 2014