Identification and Estimation of the Probabilities of Potential Outcome Types Using Covariate Information in Studies with Non-compliance

Yuta Kawakami, Ryusei Shingaki, Manabu Kuroki

Department of Mathematics, Physics, Electrical Engineering and Computer Science
Graduate School of Engineering Science, Yokohama National University
79-5 Tokiwadai, Hodogaya-ku, Yokohama 240-8501 JAPAN
kawakami-yuta-yd@ynu.jp

Abstract

We propose novel identification conditions and a statistical estimation method for the probabilities of potential outcome types using covariate information in randomized trials in which the treatment assignment is randomized but subject compliance is not perfect. Different from existing studies, the proposed identification conditions do not require strict assumptions such as the assumption of monotonicity. When the probabilities of potential outcome types are identifiable through the proposed conditions, the problem of estimating the probabilities of potential outcome types is reduced to that of singular models. Thus, the probabilities cannot be evaluated using standard statistical likelihood-based estimation methods. Rather, the proposed identification conditions show that we can derive consistent estimators of the probabilities of potential outcome types via the method of moments, which leads to the asymptotic normality of the proposed estimators through the delta method under regular conditions. We also propose a new statistical estimation method based on the bounded constrained augmented Lagrangian method to derive more efficient estimators than can be derived through the method of moments.

Introduction

Practical Background

A central concern in practical sciences is to elucidate the cause-effect relationship between a treatment and an outcome. Randomized trials have been regarded as a more reliable and powerful tool with which to evaluate causal effects than observational studies, where confounding, information bias, and selection bias all hinder the evaluation of causal effects from observed data. In contrast, to evaluate the likelihood that one event will cause another event, it is necessary to simultaneously derive the results for the same subjects receiving experimental and control treatments. However, doing so is difficult, even in successful randomized trials.

The importance of this topic can be seen in various fields of the practical sciences. For example, consider the following statements regarding the randomized clinical trial (RCT) comparing the coronavirus disease 2019 (COVID-19) vaccine with a placebo, as reported by Cohen (2020):

> "An EUA (emergency use authorization) for a vaccine might also make it more difficult to recruit people for clinical trials of that vaccine and others because participants might not want to take the risk of receiving a placebo when they can get a shot of a product that's authorized for use."
>
> "Vaccines go into healthy people, so putting them into use before fully assessing their risks and benefits is a bigger gamble than issuing an EUA for an experimental treatment for someone already ill."

Cohen's report implies that there are two issues in RCTs in terms of evaluating the causal effect of the COVID-19 vaccine. First, there may be subjects who do not comply with their treatment assignment. In such a situation, the RCT may fail to evaluate the causal effect, even if it successfully recruits a sufficient number of subjects. To derive the relevant results in this situation, it is important to classify the subject's situation into four types: "compliance," "defiance," "always take," and "never take." Here, "compliance" denotes a situation in which subjects receive the vaccine if and only if they are assigned to the vaccination group. "Defiance" denotes a situation in which subjects do the opposite of their treatment assignment status. "Always-take" denotes a situation in which subjects always receive the vaccine even if they are assigned to the placebo group. "Never-take" denotes a situation in which subjects do not receive the vaccine even if they are assigned to the vaccination group. Here, "defiance," "always-take," and "never-take" are collectively called "non-compliance" or "imperfect compliance." Second, it is uncertain whether the severity of the disease would decrease if unvaccinated healthy subjects are (counterfactually) vaccinated. To administer the vaccine to subjects who are most likely to benefit from it, classification of the subject's situation into four types – "causative," "preventive," "doomed," and "immune" – is useful. Here, "causative" denotes a situation in which the disease occurs if and only if the subject receives the placebo. "Preventive" denotes a situation in which the disease occurs if and only if the subject receives the vaccine. "Doomed" denotes a situation in which the treatment received is irrelevant in the sense that the disease occurs regardless of whether the vaccine or the placebo is received. "Immune" denotes a situation in which the treatment received is irrele-

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

vant in the sense that the disease does not occur regardless of whether the vaccine or the placebo is received.

For an effective vaccination policy, it is better to target subjects in the "causative" group since the vaccination can decrease the severity of the disease for these subjects, but otherwise not. In contrast, the probability of the "preventive" situation is useful for evaluating the severity of the vaccination since non-vaccination decreases the severity of the disease for these subjects, but otherwise not.

The above types are called "potential outcome types," as each subject belongs to one of $4 \times 4 = 16$ types but we do not know from the observed data to which type they belong. The central aim of this paper is to evaluate "probabilities of 16 potential outcome types" in randomized trials with non-compliance. The probabilities of these potential outcome types are fundamental components of causal inference since, if they are identifiable, the causal effect is also identifiable, but not vice versa (Pearl 1999, 2009).

Theoretical Background

The identification and estimation of the probabilities of potential outcome types are important issues in causal inference. One representative example of the probabilities of potential outcome types involves the probabilities of causation that enable us to evaluate the probabilistic aspects of "necessity cause," "sufficiency cause," and "necessity and sufficiency cause." The probabilities of causation have been widely utilized in such areas as epidemiology, risk analysis, tort law liability, social program impact evaluation, and traffic conflict (Lagakos and Mosteller 1986; Khoury et al. 1989; Heckman, Smith, and Clements 1997; Beyea and Greenland 1999; Cai and Kuroki 2005; Yamada and Kuroki 2019). Currently, they are among the fundamental concepts of successful explanation in the field of explainable artificial intelligence (XAI) (Galhotra, Pradhan, and Salimi 2021; Mothilal et al. 2020; Watson et al. 2021). Thus, the evaluation of the probabilities of potential outcome types has become an important issue in terms of extending the range of solvable identification and estimation problems in causal inference.

Pearl (1999, 2009) developed formal semantics for the probabilities of causation based on structural causal models. The probabilities of causation are formulated based on those of potential outcome types. Since one cannot simultaneously observe the results of the same subjects receiving the experimental treatment and the control treatment in reality, these quantities are not identifiable, even in successful randomized trials (Pearl 2009). To solve this problem, Tian and Pearl (2000) demonstrated how to bound these quantities from data obtained in experimental and observational studies. Tian and Pearl's bounds provide the range within which the probabilities of causation must lie. Subsequently, Kuroki and Cai (2011) derived narrower bounds of the probabilities of causation than Tian and Pearl's bounds using covariate information. Dawid, Murtas, and Musio (2016) and Mueller, Li, and Pearl (2021) provided the bounds of the probabilities of causation using mediator variables. However, it has been pointed out that these bounds are too wide to effectively evaluate the probabilities of causation.

To overcome this difficulty, Tian and Pearl (2000) noted that the probabilities of causation are identifiable if monotonicity (e.g., no-prevention) can be assumed and the causal effects are identifiable, and Pearl (2009) showed that specific functional relationships between cause and effect lead to the identification of the probabilities of causation. In addition, in the context of natural direct and indirect effects (Pearl 2001), under the assumption of there being no unmeasured confounding, Robins and Richardson (2011) stated that the probabilities of potential outcome types are identifiable (i) if two potential outcome variables are independent, or (ii) if one potential outcome variable can be deterministically formulated as a function of the other potential outcome variable. These prior studies showed that the probabilities of potential outcome types play an important role in solving various problems in causal inference. However, there has been much less discussion of identifying the probabilities of potential outcome types when the present assumptions are violated.

This paper proposes novel identification conditions, which are developments of Kawakami (2021), and a statistical estimation method for estimating the probabilities of potential outcome types in randomized trials in which the treatment assignment is randomized but subject compliance is imperfect. The proposed identification conditions enable us to derive consistent estimators of the probabilities of potential outcome types without relying on the previously used assumptions. Notably, the proposed conditions achieve (i) a major expansion of the existing results to evaluate the probabilities of potential outcome types, from "4 potential outcome types" to "16 potential outcome types", and (ii) the identification of the probabilities of "compliance,"' "defiance," "always take," and "never take," which are important measures in medical science (Frangakis and Rubin 2002; Kowalski 2020). In particular, although it has been pointed out that adjustment by a proxy variable of the outcome (graphically, the descendant of the outcome) for the treatment provides a biased estimate of the causal effect, this paper shows that the use of such a proxy variable makes the causal effect identifiable, at least in some situations. In addition, different from existing studies, the proposed identification conditions do not require strict assumptions such as monotonicity (Pearl 2009). When the probabilities of potential outcome types are identifiable through the proposed conditions, the problem of estimating the probabilities of potential outcome types is reduced to that of singular models. Thus, these probabilities cannot be evaluated by standard likelihood-based estimation methods. Rather, the proposed identification conditions show that we can derive consistent estimators of the probabilities of potential outcome types via the method of moments, which leads to the asymptotic normality of the proposed estimators through the delta method under regular conditions. Noting that the method of moments may not provide efficient estimators, we also propose a new statistical estimation method based on the bounded constrained augmented Lagrangian method (Birgin and Martínez 2020) to derive more efficient estimators than can be derived via the method of moments.

Due to space constraints, proofs, some simulation experiments, and details of a case study are provided in the Supplementary Material.

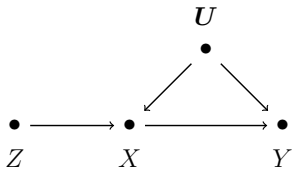


Figure 1: Graphical representation of a randomized trial with non-compliance

Problem Description

To describe our problem, let us consider a randomized trial with the purpose of evaluating the causal effect of an experimental treatment (e.g., the COVID-19 vaccine) in comparison with a control treatment (e.g., a placebo), as shown in Figure 1. For the graph-theoretic terminology and basic theory of the structural causal models used in this paper, we refer readers to Pearl (2009). In addition, in this paper, we assume that readers are familiar with the basic theory of causal inference (e.g., Imbens and Rubin 2015; Pearl 2009).

Intuitively, in Figure 1 with variables X, Y, and Z and a set $\boldsymbol{U}$ of variables, a directed edge from X to Y indicates that X could have an effect on Y. In addition, the absence of a directed edge from Y to X indicates that Y cannot be a cause of X. A directed edge from $\boldsymbol{U}$ to Y indicates that some elements of $\boldsymbol{U}$ could affect Y. Furthermore, the absence of a directed edge from Y to $\boldsymbol{U}$ indicates that Y cannot be a cause of any element of $\boldsymbol{U}$. A directed path from Z to Y through X indicates that Z could affect Y through X. X is also called a mediator variable regarding the directed path from Z to Y; X could be affected by Z and have an effect on Y. A similar interpretation is applied to other directed edges and paths. This situation often appears in the medical and statistical literature, for example, Multiple Risk Factor Intervention Trial Research Group (1982), Balke and Pearl (1997), Cai, Kuroki, and Sato (2007), and Lui (2011).

In this situation, we assume that Z, X, and Y are observed dichotomous variables, where Z represents the randomized treatment assignment, X represents the treatment actually received, and Y represents the outcome. In addition, let z, x, and y represent the values taken by the variables Z, X, and Y, respectively, with the following meanings: for $z \in \{z_0, z_1\}$, z_1 indicates subjects randomized to the experimental treatment, while z_0 indicates subjects randomized to the control treatment; for $x \in \{x_0, x_1\}$, x_1 indicates receiving the experimental treatment, while x_0 indicates receiving the control treatment; and for $y \in \{y_0, y_1\}$, y_1 indicates non-disease, while y_0 indicates disease. $\boldsymbol{U}$ represents the set of all discrete and continuous variables, both observed and unobserved, that are not affected by X or Y. A variable that is not affected by a treatment X is called a covariate.

In this situation, randomized treatment assignment Z satisfies the instrumental variable (IV) assumptions: (i) Z is associated with X, (ii) Z is independent of $\boldsymbol{U}$, and (iii) Z is conditionally independent of Y given $\boldsymbol{U}$ and $\{X\}$ (Greenland 2000). Then, Z is called an IV relative to (X, Y). Here, regarding X, Y, and Z, it is straightforward to extend our results from the case of dichotomous observed variables to the case of multivalued observed variables. In particular, as Balke and Pearl (1997) stated, a multivalued or continuous outcome can be accommodated in the model using event $Y < y$ as an outcome variable. For a related discussion, refer to Galhotra, Pradhan, and Salimi (2021) and Kada, Cai, and Kuroki (2013). In addition, when the treatment is continuous, according to Balke and Pearl (1997), in some situations, it is reasonable to assume that there exists a treatment interval around each x, within which a subject's outcome is homogeneous. Then, it is possible to apply our idea.

Notation and Idea

Let N denote the sample size. For $x \in \{x_0, x_1\}$ and $y \in \{y_0, y_1\}$, let $\mathrm{pr}(X = x, Y = y) = \mathrm{pr}(x, y)$ be the joint probability of $(X, Y) = (x, y)$, $\mathrm{pr}(Y = y \,|\, X = x) = \mathrm{pr}(y \,|\, x)$ be the conditional probability of $Y = y$ given $X = x$, and $\mathrm{pr}(X = x) = \mathrm{pr}(x)$ be the marginal probability of $X = x$. Similar notation is used for other probabilities. Then, in principle, for $x \in \{x_0, x_1\}$, the i-th of the N subjects has a potential outcome variable $Y_x(i)$ that would have resulted if X had been x for the i-th subject. Here, $Y_x(i) = y$ denotes that "Y takes the value y when X is experimentally set to x for the i-th subject" or the counterfactual sentence that "Y would be y had X been x for the i-th subject." The potential outcome variable $Y_x(i)$ is observed only if X is x for the i-th subject, denoted as $X(i) = x$. This property is referred to as consistency (Robins 1989; Pearl 2009), and formulated as

$$X(i) = x \implies Y_x(i) = Y \tag{1}$$

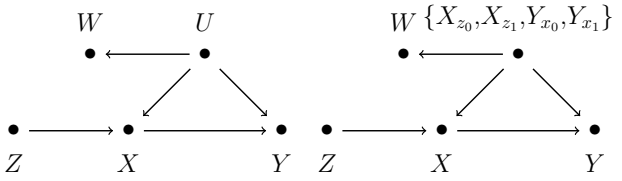
for the i-th subject.

This paper assumes the stable unit treatment value assumption (Imbens and Rubin 2015), which can be summarized as follows: (i) the treatment status of any subject does not affect the outcomes of the other subjects (no interference) and (ii) the treatments of all subjects are comparable (no variations in treatment). Thus, when N subjects in the study are considered random samples from the population of interest, $Y_x(i)$ is referred to as the value of a random variable, Y_x.

The causal effect of $X = x$ on $Y = y$ is defined as a contrast of two causal risks $\mathrm{pr}(Y_x = y)$ and $\mathrm{pr}(Y_{x'} = y)$. According to Pearl (2009), the causal risk of $X = x$ on $Y = y$ is represented as

$$\mathrm{pr}(Y_x = y) = \sum_u \mathrm{pr}(y|x, \boldsymbol{u})\, \mathrm{pr}(\boldsymbol{u}) \tag{2}$$

based on a set $\boldsymbol{U}$ of background variables. Here, summation signs are replaced by integrals whenever the summed variables are continuous. Equation (2) is identifiable and is given by the formula $\mathrm{pr}(Y_x = y) = \mathrm{pr}(y|x)$, if an ideal randomized trial with X is feasible. Here, "identifiable" means that the causal quantities, such as $\mathrm{pr}(Y_x = y)$, can be estimated consistently from a joint probability of observed variables. In contrast, when it is difficult to conduct a randomized trial and only observed data are available, we can still evaluate the causal effects according to the conditionally ignorable treatment assignment condition (Rosenbaum and Rubin 1983) or, graphically, the back-door criterion (Pearl 2009). In other words, for treatment X, if there exists such a set $\boldsymbol{S}$ of observed covariates that X is conditionally independent of (Y_{x_0}, Y_{x_1}) given $\boldsymbol{S}$, then we can say that treatment



(a) Unobserved covariates (b) Potential outcome types

Figure 2: Graphical representation of covariate information in a randomized trial with non-compliance

assignment is conditionally ignorable given $\boldsymbol{S}$. In this case, the causal risk is identifiable and are given by the formula

$$\mathrm{pr}(Y_x = y) = \sum_s \mathrm{pr}(y \mid x, \boldsymbol{s})\,\mathrm{pr}(\boldsymbol{s}). \tag{3}$$

Although there are other identification conditions of causal effects (e.g., Tian and Pearl 2002; Pearl 2009), the present paper does not cover them due to space constraints.

When compliance is not perfect and a set of observed variables is insufficient for identification, the causal effect is not identifiable without any further assumption. To solve this problem, by using the data of the randomized treatment assignment, the treatment actually received, and the outcome from a randomized trial, Balke and Pearl (1997) provided the sharp bounds of causal effects under the IV assumptions. Balke and Pearl's bounds provide the range within which the causal effect must lie. In addition, noting that a set $\boldsymbol{U}$ of covariates includes observed variables, Cai, Kuroki, and Sato (2007) derived narrower bounds than Balke and Pearl's bounds using the observed covariate information. According to Cai, Kuroki, and Sato (2007), such covariate information does not need to include confounding factors to narrow the bounds. This observation regarding covariate information motivates us to evaluate "probabilities of 16 potential outcome types" in studies with non-compliance.

Identification

Referring to the problem described in the previous section, we formulate the problem of evaluating the probabilities of potential outcome types based on the directed graph shown in Figure 2 (a). Here, a covariate W, in Figure 2 (a), is measured as a proxy variable of a set $\boldsymbol{U}$ of covariates. Note that W can be a set of discrete and/or continuous variables. Then, Figure 2 (a) graphically represents the data-generating process,

$$\begin{aligned} Y = f_y(X, \boldsymbol{U}, \epsilon_y), \quad X = f_x(\boldsymbol{U}, Z, \epsilon_x), \\ W = f_w(\boldsymbol{U}, \epsilon_w), \quad Z = f_z(\epsilon_z), \end{aligned} \tag{4}$$

where ϵ_x, ϵ_y, ϵ_z, and ϵ_w are independent random disturbances and are also independent of $\boldsymbol{U}$. When structural equation models, such as equation (4), are used to represent the data-generating process, the corresponding graph shown in Figure 2 (a) is called a causal diagram.

Regarding Figure 2 (a), note that this paper considers a situation where $\boldsymbol{U}$ can include the uncertain number of all discrete and continuous covariates that influence the way a subject responds to treatments. Accordingly, in many situations, it is reasonable to assume the existence of a covariate, W, that is conditionally independent of $\{X, Y, Z\}$ given $\boldsymbol{U}$ and, thus, it would not be difficult to observe such a proxy variable that satisfies the condition. Then, irrespective of the complexity of $\boldsymbol{U} \cup \{\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_w\}$, the impact of $\boldsymbol{U} \cup \{\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_w\}$ on Y remains restricted to the modification of the functional relationship between X and Y. This yields four functions for two dichotomous variables, X and Y, and thus, the value taken by $\boldsymbol{U} \cup \{\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_w\}$ selects one of these four functions (Pearl 2009). Considering these observations, the states of $\boldsymbol{U} \cup \{\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_w\}$ are divided into the following four potential outcome types: "doomed" u_1, "causative" u_2, "preventive" u_3, and "immune" u_4, i.e.,

$$\begin{aligned} u_1 = (Y_{x_0} = y_0, Y_{x_1} = y_0), \quad u_2 = (Y_{x_0} = y_0, Y_{x_1} = y_1), \\ u_3 = (Y_{x_0} = y_1, Y_{x_1} = y_0), \quad u_4 = (Y_{x_0} = y_1, Y_{x_1} = y_1). \end{aligned}$$

According to this partition of the states of $\boldsymbol{U} \cup \{\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_w\}$, it is redefined as U taking a value of u ($u \in \{u_1, u_2, u_3, u_4\}$). Similarly, in addition to the division of the above four potential outcome types, the states of $\boldsymbol{U} \cup \{\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_w\}$ are also divided into the following four potential outcome types: "never take" v_1, "compliance" v_2, "defiance" v_3, and "always take" v_4, i.e.,

$$\begin{aligned} v_1 = (X_{z_0} = x_0, X_{z_1} = x_0),\ v_2 = (X_{z_0} = x_0, X_{z_1} = x_1), \\ v_3 = (X_{z_0} = x_1, X_{z_1} = x_0),\ v_4 = (X_{z_0} = x_1, X_{z_1} = x_1). \end{aligned}$$

According to this partition of the states of $\boldsymbol{U} \cup \{\epsilon_x, \epsilon_y, \epsilon_z, \epsilon_w\}$, it is redefined as V taking a value of v ($v \in \{v_1, v_2, v_3, v_4\}$). Then, the corresponding probabilities that we wish to evaluate are the probabilities of potential outcome types, i.e., $\mathrm{pr}(u_i)$, $\mathrm{pr}(v_j)$, $\mathrm{pr}(u_i, v_j)$, $i, j = 1, 2, 3, 4$. These probabilities are not identifiable even in a successful randomized trial without any further information because Y_{x_1} and Y_{x_0} cannot be observed simultaneously for each subject. In this paper, we use Z as an instrumental variable (IV) to solve the identification problem, where the IV assumptions are counterfactually represented as follows: (i) X_z is a nontrivial function of z, (ii) $Y_{x,z} = Y_x$ holds for any x and z (exclusion restriction), and (iii) Z is independent of $(X_{z_0}, X_{z_1}, Y_{x_0}, Y_{x_1})$.

For any x, y, z, and w, we assume that Figure 2 (a) can be redescribed as Figure 2 (b) and that the corresponding recursive factorization of the joint probabilities of Y, Z, and W given X, $\mathrm{pr}(y, z, w|x)$, is given by

$$\begin{aligned} \mathrm{pr}(y, z, w \mid x) = \sum_{i=1}^{4} \sum_{j=1}^{4} \mathrm{pr}(y \mid x, u_i, v_j)\,\mathrm{pr}(z \mid x, u_i, v_j) \\ \times \mathrm{pr}(w \mid u_i, v_j)\,\mathrm{pr}(u_i, v_j|x) \end{aligned} \tag{5}$$

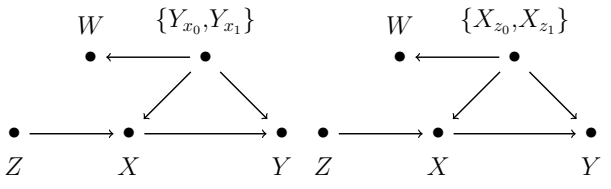
for $\mathrm{pr}(y, z, w, u, v \mid x) > 0$. When W is a variable with the number of values $k \geq 4$, say w_1, w_2, w_3, w_4, for $x \in \{x_0, x_1\}$, the block matrices are defined as

$$[P_x; Q_x] = \begin{pmatrix} P_{1,x} & Q_{1,x} \\ P_{2,x} & Q_{2,x} \end{pmatrix} \tag{6}$$

where

$$P_{1,x} = \begin{pmatrix} 1 & \mathrm{pr}(z_0 \mid x) \\ \mathrm{pr}(w_1 \mid x) & \mathrm{pr}(w_1, z_0 \mid x) \end{pmatrix}, \tag{7}$$

$$P_{2,x} = \begin{pmatrix} \mathrm{pr}(w_2 \mid x) & \mathrm{pr}(w_2, z_0 \mid x) \\ \mathrm{pr}(w_3 \mid x) & \mathrm{pr}(w_3, z_0 \mid x) \end{pmatrix}, \tag{8}$$



(a) $\{Y_{x_0}, Y_{x_1}\}$ (b) $\{X_{z_0}, X_{z_1}\}$

Figure 3: Graphical representation of potential outcome types in a randomized trial with non-compliance

$$Q_{1,x} = \begin{pmatrix} \mathrm{pr}(y_0 \mid x) & \mathrm{pr}(y_0, z_0 \mid x) \\ \mathrm{pr}(y_0, w_1 \mid x) & \mathrm{pr}(y_0, w_1, z_0 \mid x) \end{pmatrix}, \quad (9)$$

$$Q_{2,x} = \begin{pmatrix} \mathrm{pr}(y_0, w_2 \mid x) & \mathrm{pr}(y_0, w_2, z_0 \mid x) \\ \mathrm{pr}(y_0, w_3 \mid x) & \mathrm{pr}(y_0, w_3, z_0 \mid x) \end{pmatrix} \quad (10)$$

for $x \in \{x_0, x_1\}$.

First, regarding the probabilities of potential outcome types $\mathrm{pr}(u_1)$, $\mathrm{pr}(u_2)$, $\mathrm{pr}(u_3)$, and $\mathrm{pr}(u_4)$, under the situation shown in Figure 3 (a), the following theorem is derived:

Theorem 1. *Letting W be a variable taking $k(\geq 4)$ values, e.g., $w \in \{w_1, w_2, w_3, w_4\}$, the probabilities of potential outcome types $\mathrm{pr}(u_1)$, $\mathrm{pr}(u_2)$, $\mathrm{pr}(u_3)$, and $\mathrm{pr}(u_4)$ are identifiable if the following conditions are satisfied:*

- *(i) probabilities $\mathrm{pr}(y, z, w|x)$ are available for $x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_0, z_1\}$ and $w \in \{w_1, w_2, w_3, w_4\}$.*
- *(ii) for positive probabilities $\mathrm{pr}(w|x, z, u)$, $\mathrm{pr}(w|x, z, u) = \mathrm{pr}(w|u)$ holds for $x \in \{x_0, x_1\}$, $z \in \{z_0, z_1\}$, and $w \in \{w_1, w_2, w_3, w_4\}$.*
- *(iii) the block (4×4) matrix $[P_x; Q_x]$ and 2×2 matrices $P_{i,x}$, $Q_{i,x}$, and $P_{i,x} - Q_{i,x}$ are invertible for $i = 1, 2$ and $x \in \{x_0, x_1\}$.*
- *(iv) the second column vectors of $Q_{2,x}Q_{1,x}^{-1}$ and $(P_{2,x} - Q_{2,x})(P_{1,x} - Q_{1,x})^{-1}$ are different from those of $Q_{2,x'}Q_{1,x'}^{-1}$ and $(P_{2,x'} - Q_{2,x'})(P_{1,x'} - Q_{1,x'})^{-1}$ for $x \neq x'$.*
- *(v) Z is the IV relative to (X, Y): X_z is a nontrivial function of z, $Y_{x,z} = Y_x$ holds for any x and z, and Z is independent of $(X_{z_0}, X_{z_1}, Y_{x_0}, Y_{x_1})$.*

The proof of Theorem 1 is given in Supplementary Material A. Theorem 1 shows that there are some cases where causal effects are identifiable even if the representative identification conditions, e.g., the back-door criterion, are not applicable to estimating causal effects. The two causal risks are given by the formula

$$\mathrm{pr}(Y_{x_0} = y_0) = \sum_{i=1,2} \mathrm{pr}(u_i), \quad \mathrm{pr}(Y_{x_1} = y_0) = \sum_{i=1,3} \mathrm{pr}(u_i).$$

As for the probabilities of potential outcome types $\mathrm{pr}(v_1)$, $\mathrm{pr}(v_2)$, $\mathrm{pr}(v_3)$, and $\mathrm{pr}(v_4)$, when W is a variable with the number of values $k \geq 3$, that is, $w \in \{w_1, w_2, w_3\}$, we define

$$P_{x,z}^{w} = \begin{pmatrix} 1 & \mathrm{pr}(y \mid x, z) \\ \mathrm{pr}(w \mid x, z) & \mathrm{pr}(y, w \mid x, z) \end{pmatrix} \quad (11)$$

for $x \in \{x_0, x_1\}$ and $z \in \{z_0, z_1\}$. Then, under the situation shown in Figure 3 (b), we derive the following new theorem:

Theorem 2. *Letting W be a variable taking $k(\geq 3)$ values, e.g., $w \in \{w_1, w_2, w_3\}$, the probabilities of potential outcome types $\mathrm{pr}(v_1)$, $\mathrm{pr}(v_2)$, $\mathrm{pr}(v_3)$, and $\mathrm{pr}(v_4)$ are identifiable if the following conditions are satisfied:*

- *(i) probabilities $\mathrm{pr}(y, w|x, z)$ are available for $x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_0, z_1\}$, and $w \in \{w_1, w_2, w_3\}$.*
- *(ii) for positive probabilities $\mathrm{pr}(y, w|x, z, v)$, both $\mathrm{pr}(y|x, w, z, v) = \mathrm{pr}(y|x, v)$ and $\mathrm{pr}(w|x, z, v) = \mathrm{pr}(w|v)$ hold for $x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_0, z_1\}$, $w \in \{w_1, w_2, w_3\}$, and $v \in \{v_1, v_2, v_3, v_4\}$.*
- *(iii) $P_{x,z}^{w}$ are invertible for any x, y, z, and w, and*

$$\frac{\det(P_{x,z}^{w_1})}{\det(P_{x,z}^{w})} \neq \frac{\det(P_{x',z'}^{w_1})}{\det(P_{x',z'}^{w})} \quad (12)$$

  for $w \neq w_1$ and $(x, z) \neq (x', z')$ ($x, x' \in \{x_0, x_1\}$, $z, z' \in \{z_0, z_1\}$, $w \in \{w_2, w_3\}$).
- *(iv) Z is the IV relative to (X, Y).*

The proof of Theorem 2 is given in Supplementary Material B. Generally, the adjustment by the proxy variable of the outcome for the treatment provides a biased estimation for the causal effect of interest. However, Theorem 2 shows that it is remarkable that the use of the proxy outcome makes the probabilities of the potential outcome types identifiable in some situations.

Finally, the following theorem, together with Theorems 1 and 2, enables us to identify the probabilities, $\mathrm{pr}(u, v)$, of sixteen potential outcome types:

Theorem 3. *Letting W be a variable taking $k(\geq 4)$ values, e.g., $w \in \{w_1, w_2, w_3, w_4\}$, the probabilities of potential outcome types $\mathrm{pr}(u_i, v_j)$ $(i, j = 1, 2, 3, 4)$ are identifiable if the following conditions are satisfied:*

- *(i) both $\mathrm{pr}(v|w)$ and $\mathrm{pr}(u|w)$ are identifiable for $u \in \{u_1, u_2, u_3, u_4\}$, $v \in \{v_1, v_2, v_3, v_4\}$ and $w \in \{w_1, w_2, w_3, w_4\}$.*
- *(ii) one of the following conditions is satisfied:*
  - *(ii-a) for positive probabilities $\mathrm{pr}(u, w|v)$, $\mathrm{pr}(u|v, w) = \mathrm{pr}(u|v)$ holds for $u \in \{u_1, u_2, u_3, u_4\}$, $v \in \{v_1, v_2, v_3, v_4\}$ and $w \in \{w_1, w_2, w_3, w_4\}$, and the 4×4 matrix $(\mathrm{pr}(v_i|w_j))_{i,j=1,\cdots,4}$ is invertible.*
  - *(ii-b) for positive probabilities $\mathrm{pr}(v, w|u)$, $\mathrm{pr}(v|u, w) = \mathrm{pr}(v|u)$ holds for $u \in \{u_1, u_2, u_3, u_4\}$, $v \in \{v_1, v_2, v_3, v_4\}$, and $w \in \{w_1, w_2, w_3, w_4\}$, and the 4×4 matrix $(\mathrm{pr}(u_i|w_j))_{i,j=1,\cdots,4}$ is invertible.*

The proof is straightforward; for example, under conditions (i) and (ii-a), we derive

$$\begin{pmatrix} \mathrm{pr}(u|v_1) \\ \vdots \\ \mathrm{pr}(u|v_4) \end{pmatrix} = \begin{pmatrix} \mathrm{pr}(v_1|w_1) & \cdots & \mathrm{pr}(v_4|w_1) \\ \vdots & \ddots & \vdots \\ \mathrm{pr}(v_1|w_4) & \cdots & \mathrm{pr}(v_4|w_4) \end{pmatrix}^{-1} \times \begin{pmatrix} \mathrm{pr}(u|w_1) \\ \vdots \\ \mathrm{pr}(u|w_4) \end{pmatrix} \quad (13)$$

for $u \in \{u_1, u_2, u_3, u_4\}$, which shows that $\mathrm{pr}(u|v)$ is identifiable, and thus leads to $\mathrm{pr}(u, v) = \sum_{i=1}^{4} \mathrm{pr}(u|v)\,\mathrm{pr}(v|w_i)$ for $u \in \{u_1, u_2, u_3, u_4\}$ and $v \in \{v_1, v_2, v_3, v_4\}$.

Estimation

The proofs of Theorems 1 and 2 in the Supplementary Material provide the explicit formula for the probabilities of potential outcome types; thus, it is not difficult to show that the method of moments provides consistent estimators of these probabilities with asymptotic normality through the delta method under regular conditions (Ferguson 1996). For example, when the subjects in the study are considered random samples from a multinomial distribution under the conditions of Theorems 1, 2, and 3, the method of moments estimators are consistent estimators of these probabilities and asymptotically follow a normal distribution. However, since the explicit formulas for the asymptotic variances in the method of moments estimators are complicated, it is better to use a bootstrap procedure to evaluate the statistical properties of these estimators. In this section, to illustrate more efficient estimation, we refer to Theorem 1 as an example, and formulate a statistical estimation method for $\mathrm{pr}(u)$ based on the bounded constrained augmented Lagrangian method (Birgin and Martínez 2020). For details on the statistical estimation methods, the reader is referred to Supplementary Material C.1 and C.2, which address Theorems 1 and 2, respectively.

Let $\{(X_i, Y_i, Z_i, W_i)\}_{i=1}^{N}$ be a set of random samples from the data-generating process in Figure 3 (a). In addition, for $x \in \{x_0, x_1\}$, $y \in \{y_0, y_1\}$, $z \in \{z_0, z_1\}$, and $w \in \{w_1, w_2, w_3, w_4\}$, let $\widehat{\mathrm{pr}}(y, w, z|x)$ be the maximum likelihood estimators of observed probabilities $\mathrm{pr}(y, w, z|x)$. Then, we represent the plug-in estimators of P_x and Q_x as

$$\widehat{P}_x = \begin{pmatrix} \widehat{P}_{1,x} \\ \widehat{P}_{2,x} \end{pmatrix}, \quad \widehat{Q}_x = \begin{pmatrix} \widehat{Q}_{1,x} \\ \widehat{Q}_{2,x} \end{pmatrix} \tag{14}$$

for $x \in \{x_0, x_1\}$. Then, letting

$$\Theta_{x_0} = \begin{pmatrix} 1 & \theta_{11} & \theta_{12} & \theta_{13} \\ 1 & \theta_{21} & \theta_{22} & \theta_{23} \\ 1 & \theta_{31} & \theta_{32} & \theta_{33} \\ 1 & \theta_{41} & \theta_{42} & \theta_{43} \end{pmatrix}, \tag{15}$$

$$\Theta_{x_1} = \begin{pmatrix} 1 & \theta_{11} & \theta_{12} & \theta_{13} \\ 1 & \theta_{31} & \theta_{32} & \theta_{33} \\ 1 & \theta_{21} & \theta_{22} & \theta_{23} \\ 1 & \theta_{41} & \theta_{42} & \theta_{43} \end{pmatrix}, \tag{16}$$

the estimation problem of interest is formulated as the following Lagrangian function:

$$\begin{aligned} F\left(\Theta, \left\{\lambda_{i,j}^{+\,(l)}, \lambda_{i,j}^{-\,(l)} : i = 0, 1; j = 1, 2, 3, 4\right\}, \mu\right) \\ = \sum_{i=0}^{1} \|\Theta_{x_i}^T \Delta (\Theta_{x_i})^{-T} \widehat{P}_{x_i} - \widehat{Q}_{x_i}\|_F^2 \\ + \frac{\mu}{2} \left(\sum_{i=0}^{1} \sum_{j=1}^{4} \max\left\{ e_{i,j} - 1 + \frac{\lambda_{i,j}^{+}}{\mu}, 0 \right\} \right)^2 \\ + \frac{\mu}{2} \left(\sum_{i=0}^{1} \sum_{j=1}^{4} \max\left\{ -e_{i,j} + \frac{\lambda_{i,j}^{-}}{\mu}, 0 \right\} \right)^2 \\ + \sum_{i=0}^{1} \sum_{j=1}^{4} \lambda_{i,j}^{+} \max\left\{ e_{i,j} - 1 + \frac{\lambda_{i,j}^{+}}{\mu}, 0 \right\} \\ + \sum_{i=0}^{1} \sum_{j=1}^{4} \lambda_{i,j}^{-} \max\left\{ -e_{i,j} + \frac{\lambda_{i,j}^{-}}{\mu}, 0 \right\}, \end{aligned} \tag{17}$$

where $||G||_F$ is the Frobenius norm of the matrix G, and $G^{-T} = (G^T)^{-1}$. $\Theta = \{\theta_{i,j} \colon i = 1, 2, 3, 4; j = 1, 2, 3\}$ is a set of 12 parameters that are included in Θ_{x_0} and Θ_{x_1}. For $i = 0, 1$ and $j = 1, 2, 3, 4$, μ, $\lambda_{i,j}^{+}$, and $\lambda_{i,j}^{-}$ are the Lagrange multipliers. In addition, $e_{i,j}$ is the j-th element of $\Theta_{x_i}^{-T} \widehat{P}_{x_i}(1, 0)^T$ for $i = 0, 1$, and $j = 1, 2, 3, 4$, and Δ is a diagonal matrix $\Delta = \mathrm{diag}(1, 1, 0, 0)$. Once we obtain estimator $\widehat{\Theta}$ as the solution to the estimation problem, the estimators of $\{\widehat{\mathrm{pr}}(u_i|x_j) : i = 1, 2, 3, 4; j = 0, 1\}$ are given by $\widehat{\Theta}_{x_0}^{-T} \widehat{P}_{x_0} (1, 0)^T$ and $\widehat{\Theta}_{x_1}^{-T} \widehat{P}_{x_1} (1, 0)^T$. Thus, the probabilities of potential outcome types are estimated by

$$\widehat{\mathrm{pr}}(u_i) = \widehat{\mathrm{pr}}(u_i|x_0)\widehat{\mathrm{pr}}(x_0) + \widehat{\mathrm{pr}}(u_i|x_1)\widehat{\mathrm{pr}}(x_1). \tag{18}$$

(a) $\mathrm{pr}(u, v)$

	u_1	u_2	u_3	u_4
v_1	5/32	1/32	1/32	1/32
v_2	1/32	5/32	1/32	1/32
v_3	1/32	1/32	5/32	1/32
v_4	1/32	1/32	1/32	5/32

(b) $\mathrm{pr}(w|u)$ and $\mathrm{pr}(w|v)$

	u_1	u_2	u_3	u_4
	v_1	v_2	v_3	v_4
w_1	7/10	1/10	1/10	1/10
w_2	1/10	7/10	1/10	1/10
w_3	1/10	1/10	7/10	1/10
w_4	1/10	1/10	1/10	7/10

Table 1: Parameter Setting

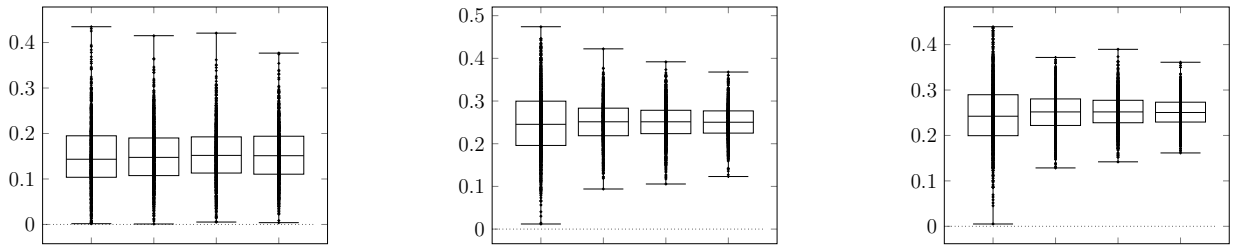
Simulation Experiments

We conducted a series of simulation experiments to examine the statistical properties of the proposed estimators, based on the bounded constrained augmented Lagrangian method (Birgin and Martínez 2020), referring to the probabilities of "causative" $\mathrm{pr}(u_2)$, "compliance" $\mathrm{pr}(v_2)$, and a combination of both "causative" and "compliance" $\mathrm{pr}(u_2, v_2)$. For simplicity, letting X, Y, Z, W, U, and V be discrete variables, we consider the causal diagrams shown in Figure 3 (a), for $\mathrm{pr}(u_2)$ and $\mathrm{pr}(u_2, v_2)$, and in Figure 3 (b), for $\mathrm{pr}(v_2)$. In Figure 3 (a), we assume $\mathrm{pr}(v|u, w) = \mathrm{pr}(v|u)$ for $u \in \{u_1, u_2, u_3, u_4\}$, $v \in \{v_1, v_2, v_3, v_4\}$, and $w \in \{w_1, w_2, w_3, w_4\}$. In addition, in order to estimate $\mathrm{pr}(v_2)$, we applied Algorithm 2 in the Supplementary Material to any combination of three distinct values from $w \in \{w_1, w_2, w_3, w_4\}$ and derived $\hat{\mathrm{pr}}(v_2)$ as the sample mean of the four estimates. The Parameter setting in this section is given in Table 1 with $\mathrm{pr}(z_1) = 0.5(\mathrm{pr}(z_0) = 1 - \mathrm{pr}(z_1))$.

Based on the joint observed probabilities $\mathrm{pr}(x, y, w|z)$ derived from the parameter setting, the statistical properties of the proposed estimators $\widehat{\mathrm{pr}}(u_2, v_2)$, $\widehat{\mathrm{pr}}(u_2)$, and $\widehat{\mathrm{pr}}(v_2)$ are verified in the simulation experiment using the parameter setting with sample sizes $N = 100$, 200, 1000, and 5000. In this situation, since $\mathrm{pr}(u_2, v_2)$ takes the value $5/32$ and both

	$\mathrm{pr}(u_2, v_2) = 0.156$				$\mathrm{pr}(u_2) = 0.250$				$\mathrm{pr}(v_2) = 0.250$			
N	100	500	1000	5000	100	500	1000	5000	100	500	1000	5000
min	0.002	0.001	0.005	0.004	0.012	0.094	0.106	0.123	0.005	0.129	0.142	0.162
1st Quantile	0.104	0.107	0.113	0.111	0.196	0.219	0.224	0.225	0.199	0.222	0.228	0.230
median	0.143	0.147	0.152	0.151	0.245	0.251	0.251	0.250	0.242	0.252	0.252	0.251
3rd Quantile	0.195	0.190	0.193	0.194	0.300	0.283	0.278	0.277	0.290	0.280	0.277	0.273
max	0.435	0.415	0.421	0.377	0.474	0.422	0.392	0.368	0.439	0.372	0.390	0.361
mean	0.150	0.152	0.154	0.156	0.248	0.250	0.250	0.251	0.247	0.251	0.253	0.251
standard error	0.072	0.061	0.061	0.059	0.077	0.048	0.044	0.038	0.069	0.042	0.037	0.032
skewness	0.560	0.481	0.364	0.370	0.110	-0.124	-0.067	0.016	0.180	-0.069	0.087	0.019
kurtosis	3.590	3.299	3.341	3.114	2.897	2.950	3.115	3.045	3.033	2.869	3.092	2.981

Table 2: Basic Statistics



(a) Boxplot of $\widehat{\mathrm{pr}}(u_2, v_2)$ (b) Boxplot of $\widehat{\mathrm{pr}}(u_2)$ (c) Boxplot of $\widehat{\mathrm{pr}}(v_2)$

Figure 4: Boxplots of the proposed estimators. The sample size is $N = 100, 500, 1000, 5000$ from left to right.

$\mathrm{pr}(u_2)$ and $\mathrm{pr}(v_2)$ take the value $1/4$, the sample means of $\widehat{\mathrm{pr}}(u_2)$ and $\widehat{\mathrm{pr}}(v_2)$ are expected to be close to $1/4$ and zero, respectively. Table 2 and Figures 4 show the basic statistics and box-and-whisker plots of $\widehat{\mathrm{pr}}(u_2, v_2)$, $\widehat{\mathrm{pr}}(u_2)$, and $\widehat{\mathrm{pr}}(v_2)$ for 1000 replications with the given sample size N, respectively. In Figures 4-6, the horizontal solid lines show the minimum/1st quantile/median/3rd quantile/maximum values of $\widehat{\mathrm{pr}}(u_2, v_2)$, $\widehat{\mathrm{pr}}(u_2)$, and $\widehat{\mathrm{pr}}(v_2)$, respectively.

From Table 2, the sample means of $\widehat{\mathrm{pr}}(u_2, v_2)$, $\widehat{\mathrm{pr}}(u_2)$, and $\widehat{\mathrm{pr}}(v_2)$ are close to the true values, and the sample standard errors become smaller as the sample size becomes larger. Thus, it seems that the proposed estimation method provides the consistent estimators of $\mathrm{pr}(u_2, v_2)$, $\mathrm{pr}(u_2)$, and $\mathrm{pr}(v_2)$. In addition, from Figures 5 and 6, the interquantile ranges for $\widehat{\mathrm{pr}}(u_2)$ and $\widehat{\mathrm{pr}}(v_2)$ are shown to be narrower and still include the true values even if the sample size is large. In contrast, the interquantile ranges for $\widehat{\mathrm{pr}}(u_2, v_2)$ include the true values but do not become so narrow even if the sample size is large. Note that Theorems 1-3 require the calculation of the inverse matrix of probability matrices with small nonzero determinants to estimate the probabilities of potential outcome types. This may be a reason why there are many outliers of the estimates for each sample size. Such outliers occur when it is difficult to judge that the proposed identification conditions hold from the observed data. Finally, it appears that both $\widehat{\mathrm{pr}}(u_2)$ and $\widehat{\mathrm{pr}}(v_2)$ are symmetrically distributed, and thus, asymptotic normality holds, but $\widehat{\mathrm{pr}}(u_2, v_2)$ may not. This finding may be due to Theorem 3, together with Theorems 1 and 2, requiring the calculation of the inverse matrix of the probabilities more than Theorems 1 and 2 and thus becoming unstable for a small sample size. In fact, in Table 2, regarding $\widehat{\mathrm{pr}}(u_2, v_2)$, both the skewness and kurtosis of the distribution of $\widehat{\mathrm{pr}}(u_2, v_2)$ become close to the skewness and kurtosis of the normal distribution as the sample size increases. For a further discussion of the simulation experiments and case study, see the Supplementary Material.

Discussion

We have proposed novel identification conditions for the probabilities of potential outcome types based on the IV assumptions with covariate information in randomized trials in which the treatment assignment is randomized but subject compliance is not perfect. When the probabilities of potential outcome types are identifiable through the proposed conditions, they cannot be evaluated by standard statistical likelihood-based estimation methods, as our estimation problem leads to that of singular models. To solve this problem, we provided consistent estimators of the probabilities of potential outcome types based on the method of moments, which leads to the asymptotic normality of the proposed estimators through the delta method under regular conditions. However, the method of moments estimator may not be efficient. To improve efficiency, we have proposed a bounded constrained augmented Lagrangian method (Birgin and Martínez 2020) to derive consistent estimators more efficiently than can the method of moments. Although the asymptotic normality of the augmented Lagrangian method in causal inference is discussed in Shingaki and Kuroki (2021), it is necessary to develop a more efficient estimation method based on singular models. In addition, we have assumed that the observed variables of interest are dichotomous in this paper. This assumption can be relaxed by allowing these variables of interest to have more than two categories; in such a situation, we can easily extend our results to multicategorical variables, which makes them applicable to a wider variety of situations. However, in such cases, it may be difficult to obtain reliable statistics on the recovered probabilities due to data sparseness. We leave this topic to future work.

Acknowledgements

The authors also thank the anonymous reviewers for their time and thoughtful comments. This work was supported by JSPS KAKENHI Grant Number 22J21928.

References

Balke, A.; and Pearl, J. 1997. Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.*, 92(439): 1171–1176.

Beyea, J.; and Greenland, S. 1999. The importance of specifying the underlying biologic model in estimating the probability of causation. *Health Physics*, 76(3): 269–274.

Birgin, E. G.; and Martínez, J. M. 2020. Complexity and performance of an augmented Lagrangian algorithm. *Optimization Methods and Software*, 35(5): 885–920.

Cai, Z.; and Kuroki, M. 2005. Variance estimators for three "probabilities of causation". *Risk Analysis*, 25(6): 1611–1620.

Cai, Z.; Kuroki, M.; and Sato, T. 2007. Non-parametric bounds on treatment effects with non-compliance by covariate adjustment. *Statist. Med.*, 26(16): 3188–3204.

Cohen, J. 2020. Here's how the US could release a COVID-19 vaccine before the election-and why that scares some. *Science*.

Dawid, A. P.; Murtas, R.; and Musio, M. 2016. Bounding the probability of causation in mediation analysis. In *Topics on Methodological and Applied Statistical Inference*, 75–84. Springer.

Ferguson, T. S. 1996. *A Course in Large Sample Theory*. Routledge.

Frangakis, C. E.; and Rubin, D. B. 2002. Principal stratification in causal inference. *Biometrics*, 58(1): 21–29.

Galhotra, S.; Pradhan, R.; and Salimi, B. 2021. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, 577–590.

Greenland, S. 2000. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*, 29(4): 722–729.

Heckman, J. J.; Smith, J.; and Clements, N. 1997. Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64(4): 487–535.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kada, A.; Cai, Z.; and Kuroki, M. 2013. Medical diagnostic test based on the potential test result approach: bounds and identification. *Journal of Applied Statistics*, 40(8): 1659–1672.

Kawakami, Y. 2021. Instrumental Variable-based Identification for Causal Effects using Covariate Information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13): 12131–12138.

Khoury, M. J.; Flanders, W. D.; Greenland, S.; and Adams, M. J. 1989. On the measurement of susceptibility in epidemiologic studies. *American Journal of Epidemiology*, 129(1): 183–190.

Kowalski, A. 2020. Counting defiers: Examples from health care. *arXiv preprint arXiv:1912.06739*.

Kuroki, M.; and Cai, Z. 2011. Statistical analysis of 'probabilities of causation' using covariate information. *Scand. J. Statist.*, 38(3): 564–577.

Lagakos, S. W.; and Mosteller, F. 1986. Assigned shares in compensation for radiation-related cancers. *Risk Analysis*, 6(3): 345–357.

Lui, K.-J. 2011. *Binary Data Analysis of Randomized Clinical Trials with Noncompliance*. John Wiley & Sons.

Mothilal, R. K.; Mahajan, D.; Tan, C.; and Sharma, A. 2020. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. *arXiv preprint arXiv:2011.04917*.

Mueller, S.; Li, A.; and Pearl, J. 2021. Causes of effects: Learning individual responses from population data. Technical Report R-505, Department of Computer Science, University of California, Los Angeles, CA.

Multiple Risk Factor Intervention Trial Research Group. 1982. Multiple risk factor intervention trial: Risk factor changes and mortality results. *Journal of the American Medical Association*, 248: 1465–1477.

Pearl, J. 1999. Probabilities Of causation: Three counterfactual interpretations and their identification. *Synthese*, 121(1): 93–149.

Pearl, J. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

Robins, J. 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*, 113–159. U.S. Public Health Service, Washington, DC.

Robins, J.; and Richardson, T. S. 2011. Alternative graphical causal models and the identification of direct effects. In *Causality and psychopathology: Finding the determinants of disorders and their cures*, 103–158. Oxford University Press.

Rosenbaum, P. R.; and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55.

Shingaki, R.; and Kuroki, M. 2021. Identification and estimation of joint probabilities of potential outcomes in observational studies with covariate information. In *Proceedings of the Thirty-Fifth Conference on Neural Information Processing Systems*, 26475–26486.

Tian, J.; and Pearl, J. 2000. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1): 287–313.

Tian, J.; and Pearl, J. 2002. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, 567–573.

Watson, D.; Gultchin, L.; Taly, A.; and Floridi, L. 2021. Local explanations via necessity and sufficiency: Unifying theory and practice. *arXiv preprint arXiv:2103.14651*.

Yamada, K.; and Kuroki, M. 2019. New traffic conflict measure based on a potential outcome model. *Journal of Causal Inference*, 7(1).