Causal inference in case-control studies

Sung Jae Jun
Sokbae Lee

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper
CWP19/20

Economic and Social Research Council

Causal Inference in Case-Control Studies*

Sung Jae Jun† and Sokbae Lee‡

Penn State Univ. Columbia Univ. and IFS

April 17, 2020

Abstract. We investigate identification of causal parameters in case-control and related studies. The odds ratio in the sample is our main estimand of interest and we articulate its relationship with causal parameters under various scenarios. It turns out that the odds ratio is generally a sharp upper bound for counterfactual relative risk under some monotonicity assumptions, without resorting to strong ignorability, nor to the rare-disease assumption. Further, we propose semparametrically efficient, easy-to-implement, machine-learning-friendly estimators of the aggregated (log) odds ratio by exploiting an explicit form of the efficient influence function. Using our new estimators, we develop methods for causal inference and illustrate the usefulness of our methods by a real-data example.

Key Words: relative risk, causality, monotonicity, case-control sample, machine learning, partial identification, semiparametric efficiency bound

JEL Classification Codes: C21, C55, C83

*We would like to thank Chuck Manski and seminar participants at Cemmap, Oxford, and Penn State for helpful comments. This work was supported in part by the European Research Council (ERC-2014-CoG-646917-ROMIA) and by the UK Economic and Social Research Council (ESRC) through research grant (ES/P008909/1) to the CeMMAP.

†Department of Economics, Penn State University, 619 Kern Graduate Building, University Park, PA 16802, suj14@psu.edu

‡Department of Economics, Columbia University, 1022 International Affairs Building, 420 West 118th Street, New York, NY 10027, sl3841@columbia.edu

1. INTRODUCTION

Empirical researchers often find it useful to work with outcome-based or case-control samples when they study rare events: cancer (Breslow and Day, 1980), infant death (Currie and Neidell, 2005), consumer bankruptcy (Domowitz and Sartain, 1999), and drug trafficking (Carvalho and Soares, 2016), among many others. Case-control sampling arises frequently in biostatistics when doctors or epidemiologists study risk factors for a rare disease: random sampling may yield only a few observations with the disease among several thousands of data. In econometrics, it is often referred to as choice-based or response-based sampling because the outcome of interest is discrete choice in many economic applications (see, e.g., Chapter 6 of Manski, 2009).

Inference methods that work with random samples are generally not suitable when data are outcome-based. In the econometrics literature, parametric estimation with outcome-based samples has been investigated by Manski and Lerman (1977), Cosslett (1981), Manski and McFadden (1981), Hsieh, Manski, and McFadden (1985), Imbens (1992), and Lancaster and Imbens (1996), among others. This strand of the literature has focused mainly on the consistency or efficiency of parametric estimators in discrete response models; see e.g. McFadden (2015) for a review. In the biostatistics and epidemiology literature (e.g. Breslow, 1996), logistic regression has been the standard workhorse model in analyzing case-control studies with a more emphasis on sampling designs.

To motivate the setup of this paper, we start with a simple example. Table 1 summarizes data from American Community Survey (ACS) 2018, cross-tabulating the likelihood of top income by educational attainment. The sample is restricted to white males residing in California with at least a bachelor's degree.[1] The binary

[1] It is extracted from IPUMS USA (Ruggles, Flood, Goeken, Grover, Meyer, Pacas, and Sobek, 2019). The ACS is an ongoing annual survey by the US Census Bureau that provides key information about US population. The IPUMS database contains samples from the 2000-2018 ACS. The ACS sample is not a case-control sample but we will use it to illustrate our proposed methods.

outcome 'Top Income' (Y) is defined to be one if a respondent's annual total pre-tax wage and salary income is top-coded.[2] The binary treatment (T) is defined to be one if a respondent has a master's degree, a professional degree, or a doctoral degree.

TABLE 1. Top Income and Education

	Beyond Bachelor's		Total
Top Income	$T = 0$	$T = 1$	
$Y = 0$	10,533	6,362	16,895
$Y = 1$	397	524	921
Total	10,930	6,886	17,816

From Table 1, the proportions of top income earners are $\mathbb{P}(Y = 1|T = 1) \approx 0.08$ and $\mathbb{P}(Y = 1|T = 0) \approx 0.04$ by educational attainment. Thus, their difference and ratio are $\mathbb{P}(Y = 1|T = 1) - \mathbb{P}(Y = 1|T = 0) \approx 0.04$ and $\mathbb{P}(Y = 1|T = 1)/\mathbb{P}(Y = 1|T = 0) \approx 2.10$, respectively. In other words, going beyond a bachelor's degree is associated with a 4% increase in the likelihood of earning top incomes and doubles the chance of earning top incomes. The corresponding odds ratio is

$$\frac{\mathbb{P}(Y = 1|T = 1)\mathbb{P}(Y = 0|T = 0)}{\mathbb{P}(Y = 0|T = 1)\mathbb{P}(Y = 1|T = 0)} \approx 2.19. \tag{1}$$

If we now look at the table in the *retrospective* manner, the proportions of going beyond a bachelor's degree are $\mathbb{P}(T = 1|Y = 1) \approx 0.57$ and $\mathbb{P}(T = 1|Y = 0) \approx 0.38$ by top income status. If we compute the odds ratio using these two probabilities, we have

$$\frac{\mathbb{P}(T = 1|Y = 1)\mathbb{P}(T = 0|Y = 0)}{\mathbb{P}(T = 0|Y = 1)\mathbb{P}(T = 1|Y = 0)} \approx 2.19, \tag{2}$$

[2]In ACS 2018, the threshold income for top-coding is different across states. In our sample extract, the top-coded income bracket has median income \$565,000 and the next highest income that is not top-coded is \$327,000.

which is the same as before. This is not a coincidence but a consequence of the Bayes rule, which is known as the invariance property of the odds ratio (e.g. Cornfield, 1951).

With case-control samples in hand, one cannot estimate the prospective odd ratio in (1) but the retrospective one in (2) is easy to obtain. The fact that the odds ratio is the same between (1) and (2)—but the difference (or the ratio) is not—explains why the odds ratio has attracted much attention in case-control studies (Breslow, 1996). Despite this invariance property, the odds ratio is not as perspicuous as the difference or ratio of two probabilities. The interpretation of the odds ratio becomes easier if it is put together with the so-called rare-disease assumption. In particular, if the outcome of interest is rare, the odds ratio approximates the ratio of two probabilities. In our example, the odds ratio of 2.19 is pretty close to the probability ratio of 2.10 since top income is a rare event (that is, $\mathbb{P}(Y = 1) \approx 0.05$).

However, learning the ratio of two probabilities is generally not sufficient to draw causal inferences, though it is useful to find association. In a lesser-known paper, Holland and Rubin (1988) adopt the potential outcome framework to illustrate how to identify the causal parameters in case-control studies: they emphasize the role of covariates and develop their analysis based on the assumption of strong ignorability. Their work is the starting point of this paper. We share their motivation and aim at surpassing their identification analysis by considering a case when strong ignorability does not hold.

Specifically, we adopt the assumptions of monotone treatment response (Manski, 1997, MTR hereafter) and monotone treatment selection (Manski and Pepper, 2000, MTS hereafter) and show that those assumptions enable us to bound the causal parameters in a meaningful way. For instance, in our example, instead of assuming strong ignorability, it might be more realistic to assume that higher degrees do not harm the chance of earning top incomes (i.e. MTR) and respondents who select higher levels of degrees are no less likely to earn top incomes, if they are randomly assigned to a different education status, than those who choose a bachelor's degree only (i.e. MTS). The MTR and MTS assumptions and related notions of monotonicity have been used in e.g. Bhattacharya, Shaikh, and Vytlacil (2008, 2012), Kreider, Pepper, Gundersen, and Jolliffe (2012), Okumura and Usui (2014), Kim, Kwon, Kwon, and Lee (2018), Machado, Shaikh, and Vytlacil (2019), and Jun and Lee (2019) among others.

To highlight the contributions of this paper, we introduce a tripartite collection of random variables: $\big(Y^*(1), Y^*(0), T^*, X^*\big)$, (Y^*, T^*, X^*), and (Y, T, X), where $Y^*(t)$ is the potential binary outcome under treatment $t \in \{0,1\}$, $Y^* = Y^*(1)T^* + Y^*(0)(1 - T^*)$, T^* and X^* are the outcome, treatment and covariates that would have been observed under random sampling, and Y, T and X are the variables that are actually observed in the outcome-based sample. As to the main causal parameter of interest, we focus on

$$\theta_{\mathrm{RR}}(x) := \frac{\mathbb{P}\{Y^*(1) = 1 | X^* = x\}}{\mathbb{P}\{Y^*(0) = 1 | X^* = x\}}, \tag{3}$$

which is *causal relative risk* conditional on $X = x$. To identify $\theta_{\mathrm{RR}}(x)$, we face two separate challenges: one results from the usual missing data problem of potential outcomes and the other stems from the fact that the researcher does not have access to (Y^*, T^*, X^*) but only to (Y, T, X).

Our contributions are two-fold. First, we articulate how the causal parameter is related with functionals of the distribution of (Y, T, X) under two different versions of outcome-based sampling schemes: i.e. the traditional case-control sampling and case-population sampling considered in Lancaster and Imbens (1996). It turns out that the odds ratio between Y and T conditional on $X = x$ is generally a sharp upper bound for $\theta_{\mathrm{RR}}(x)$ under the MTR and MTS assumptions. This interpretation does not require strong ignorability, nor does it the usual rare-disease assumption. Therefore, our identification analysis shows that we can provide the conventional estimand, i.e. the odds ratio in the sample, with causal interpretation from the perspective of partial identification (see, e.g., Manski, 2003, 2009; Tamer, 2010).

Second, we propose two novel estimation algorithms for the aggregated (log) odds ratio. For this purpose we obtain an explicit form of the efficient influence function, after which we construct suitable sample analogs. The first estimator we build is a plug-in sieve estimator (e.g. X. Chen, 2007) and the second one is a double/debiased machine learning (DML) estimator (e.g. Chernozhukov, Chetverikov, Dimirer, Duflo, Hansen, Newey, and Robins, 2018). The former is simpler but the latter accommodates LASSO-type or more general nonparametric estimators. Both estimators achieve the semiparametric efficiency bound (e.g. Newey, 1990, 1994) and can be easily implemented by using standard statistical packages. Using our estimators and the ACS data, we illustrate how to draw causal inferences based on our partial identification results as well as how to carry out a sensitivity analysis.

To the best of our knowledge, we are not cognizant of directly relevant papers in the literature. In fact, the recent econometrics literature on outcome-based sampling is rather sparse; however, it is an important reality that random sampling can be expensive when the outcome of interest is rare. The goal of this paper is to revamp outcome-based sampling from the perspective of modern econometrics. Our paper is the first paper that nonparametrically connects the three dots: outcome-based sampling, causal inference and partial identification. We provide a further discussion on how our paper is related to the existing literature in section 7.

The remainder of the paper is organized as follows. Section 2 presents the framework and identification results. We describe two sampling schemes, i.e. case-control sampling and case-population sampling, after which we discuss causal parameters and their identification. In section 3, we derive the semiparametric efficiency bound for our estimand, and in section 4, we propose two estimation algorithms. We analytically establish the local robustness property of one of our estimating equations, yielding an estimator that suits well for machine learning in section 4.3. Section 5 summarizes the main takeaways and discusses several inferential issues. Section 6 presents an empirical example using the ACS data. We conclude the paper by discussing the related literature and topics for future research in section 7. Appendices, along with an online supplement, include additional materials and all the proofs.

2. Framework

In this section, we describe the scheme of outcome-based sampling, define causal parameters and discuss their identification under two sets of assumptions: one with strong ignorability and the other without it.

2.1. **Bernoulli Sampling.** Let (Y^*, T^*, X^*) be the random variables that would have been observed if a researcher had collected data via random sampling from the population of interest, where Y^* is a binary outcome, T^* is a binary treatment, and X^* is a vector of covariates. We assume that a random sample of (Y^*, T^*, X^*) is unavailable and hence (Y^*, T^*, X^*) is not observed. Instead, we assume that we have a random sample of (Y, T, X), where (Y, T, X) represents the random variables that are actually observed in the sample that is drawn by the researcher's sampling design, i.e. Bernoulli sampling (e.g. Breslow, Robins, and Wellner, 2000), which we further describe below and discuss in section 7.

In Bernoulli sampling, the researcher draws a Bernoulli variable Y first from a pre-specified marginal distribution, after which she randomly draws (T, X) from $\mathcal{P}_y$ if and only if $Y = y$. Since $h_0 = \mathbb{P}(Y = 1)$ is part of the sampling scheme, we assume that it is known. If $\mathcal{P}_y$ is identical to the conditional distribution of (T^*, X^*) given on $Y^* = y$, then this is known as case-control sampling. The Bernoulli scheme allows for other possibilities. Below are the two leading cases that we focus on throughout the paper. In order to simplify our discussion, we first make a common-support assumption. Let $\mathcal{X}^*$ and $\mathcal{X}_y$ be the support of X^* and that of X given $Y = y$, respectively.

Assumption A (Common Support). *There is a common support $\mathcal{X}$ satisfying $\mathcal{X} = \mathcal{X}^* = \mathcal{X}_0 = \mathcal{X}_1$.*

Assumption A may not be trivial in some applications. For example, if Y^* represents breast cancer and we have two covariates to consider, i.e. gender and age, then the joint support of gender and age depends highly on whether to condition on $Y^* = 1$ or not; the breast cancer population consists mostly of women. However, in this case, using the gender variable for extra stratification is appropriate. That is, we restrict ourselves to the population of women and both X^* and X represents the age; X^* is the age that would have been drawn from the population of women and X is the age that is drawn from the subpopulation of women with or without breast cancer, depending on the corresponding value of Y. Throughout the paper, we are implicit about the possibility of stratification using extra covariates (different from those included in X^*).

Let $\mathcal{P}_y(t,x) = f_{X|Y}(x|y)\mathbb{P}(T = t|X = x, Y = y)$, where $f_{X|Y}$ is the probability density (or mass) function of X given $Y = y$ for $y = 0, 1$.

Design 1 (Case-Control Sampling)**.** *Suppose that for all $(t,x) \in \{0,1\} \times \mathcal{X}$ and for $y \in \{0,1\}$,*

$$f_{X|Y}(x|y) = f_{X^*|Y^*}(x|y) \text{ and } \mathbb{P}(T = t|X = x, Y = y) = \mathbb{P}(T^* = t|X^* = x, Y^* = y).$$

In other words, $\mathcal{P}_0$ is the distribution of (T^*, X^*) given $Y^* = 0$, while $\mathcal{P}_1$ is that of (T^*, X^*) given $Y^* = 1$.

Design 2 (Case-Population Sampling)**.** *Suppose that for all $(t,x) \in \{0,1\} \times \mathcal{X}$,*

$$f_{X|Y}(x|0) = f_{X^*}(x) \quad \text{and} \quad \mathbb{P}(T = t|X = x, Y = 0) = \mathbb{P}(T^* = t|X^* = x),$$
$$f_{X|Y}(x|1) = f_{X^*|Y^*}(x|1) \quad \text{and} \quad \mathbb{P}(T = t|X = x, Y = 1) = \mathbb{P}(T^* = t|X^* = x, Y^* = 1).$$

In other words, $\mathcal{P}_0$ represents the distribution of (T^*, X^*) of the entire population, while $\mathcal{P}_1$ is that of (T^*, X^*) conditional on $Y^* = 1$.

Design 1 is arguably the most popular form of case-control studies and design 2, which we call *case-population sampling*, is considered in Lancaster and Imbens (1996). The notation here distinguishes the original variables (Y^*, T^*, X^*) of interest from the sampled ones (Y, T, X); see, e.g., K. Chen (2001) and Xie, Lin, Yan, and Tang (2019) for using the same notational device. The advantage of this approach is that it becomes straightforward to apply asymptotic theory under random sampling to observations generated from (Y, T, X) because we can regard them as a collection of independent and identically distributed (i.i.d.) copies of (Y, T, X). The marginal distribution of (T, X) is identified from the data, while that of (T^*, X^*) is not. For instance, in design 1, we have $f_X(x) = f_{X^*|Y^*}(x|1)h_0 + f_{X^*|Y^*}(x|0)(1 - h_0) \neq f_{X^*}(x)$ if $h_0 \neq \mathbb{P}(Y^* = 1)$; h_0 is part of the sampling scheme, while $\mathbb{P}(Y^* = 1)$ is the true probability of the case in the population. Further, $f_{YX}(1, x) = f_{X^*|Y^*}(x|1)h_0 = f_{X^*}(x)\mathbb{P}(Y^* = 1|X^* = x)h_0/\mathbb{P}(Y^* = 1)$, which yields the likelihood function studied in e.g. Manski and Lerman (1977). We emphasize that $\mathbb{P}(Y = 1|X = x)$ does not have economic (or structural) interpretation like $\mathbb{P}(Y^* = 1|X^* = x)$, where the latter is often modeled by a rational behavior of an economic agent.

2.2. **Causal Functional Parameters.** To define causal functional parameters pertinent to outcome-based samples, let $Y^*(t) \in \{0, 1\}$ be the binary potential outcome of interest for treatment $t \in \{0, 1\}$. For example, in the context of Currie and Neidell (2005), $t = 1$ corresponds to exposure to air pollution, and $Y^*(1) = 1$ refers to counterfactual infant death when an infant is exposed to air pollution. With this notation the outcome Y^* can be written as $Y^* = T^*Y^*(1) + (1 - T^*)Y^*(0)$. The central counterfactual probabilities are $\mathbb{P}\{Y^*(1) = 1|X^* = x\}$ and $\mathbb{P}\{Y^*(0) = 1|X^* = x\}$. Conditional on $X^* = x$, one may consider the difference or ratio between the two counterfactual probabilities, which are called (conditional) attributable and relative risk in the literature (see, e.g. Manski, 2009). In this paper, we focus on the latter, namely causal relative risk $\theta_{\mathrm{RR}}(x)$ defined in equation (3). In view of the convenience of the odds ratio, as we demonstrated in Introduction, we also consider a *causal odds ratio* that is defined by

$$\theta_{\mathrm{OR}}(x) := \frac{\mathbb{P}\{Y^*(1) = 1|X^* = x\}}{\mathbb{P}\{Y^*(0) = 1|X^* = x\}} \frac{\mathbb{P}\{Y^*(0) = 0|X^* = x\}}{\mathbb{P}\{Y^*(1) = 0|X^* = x\}}.$$

2.3. **Identification under Strong Ignorability.** We begin this section by articulating how the odds ratio in the sample is related with some population quantities under each sampling design. Let $\mathrm{OR}(x)$ be the odds ratio given $X = x$ that is observed in the sample: i.e.

$$\mathrm{OR}(x) := \frac{\mathbb{P}(Y = 1|T = 1, X = x)}{\mathbb{P}(Y = 1|T = 0, X = x)} \frac{\mathbb{P}(Y = 0|T = 0, X = x)}{\mathbb{P}(Y = 0|T = 1, X = x)}, \tag{4}$$

where we assume that $0 < \mathrm{OR}(x) < \infty$ for all $x \in \mathcal{X}$ throughout the paper. Similarly, we define $\mathrm{OR}^*(x)$ and $\mathrm{RR}^*(x)$ by the conditional odds ratio and relative risk, respectively, in the population: i.e.

$$\mathrm{OR}^*(x) := \frac{\mathbb{P}(Y^* = 1|T^* = 1, X^* = x)}{\mathbb{P}(Y^* = 1|T^* = 0, X^* = x)} \frac{\mathbb{P}(Y^* = 0|T^* = 0, X^* = x)}{\mathbb{P}(Y^* = 0|T^* = 1, X^* = x)}, \tag{5}$$

$$\mathrm{RR}^*(x) := \frac{\mathbb{P}(Y^* = 1|T^* = 1, X^* = x)}{\mathbb{P}(Y^* = 1|T^* = 0, X^* = x)}. \tag{6}$$

Since we do not have a random sample of (Y^*, T^*, X^*), identification of $\mathrm{OR}^*(x)$ or $\mathrm{RR}^*(x)$ is a priori unclear. However, the Bayes rule shows the following result.

Lemma 1. *Under design 1, we have* $\mathrm{OR}(x) = \mathrm{OR}^*(x)$ *for all* $x \in \mathcal{X}$. *Similarly, under design 2, we have* $\mathrm{OR}(x) = \mathrm{RR}^*(x)$ *for all* $x \in \mathcal{X}$.

Lemma 1 shows how to relate the odds ratio in the case-control sample (respectively, case-population sample) with the odds ratio (respectively, relative risk) of the population. It requires additional assumptions to connect the odds ratio or relative risk of the population with the causal parameters defined in terms of the potential outcomes. The simplest approach is to use the idea of strong ignorability (see, e.g., Imbens and Rubin, 2015). In our context, strong ignorability consists of the following two assumptions.

Assumption B (Overlap). *For all* $(t, x) \in \{0, 1\} \times \mathcal{X}$, *we have*

$$0 < \mathbb{P}\{Y^*(t) = 1|X^* = x\} < 1 \quad \text{and} \quad 0 < \mathbb{P}(T^* = 1|X^* = x) < 1.$$

Assumption C (Unconfoundedness)**.** *For all $t \in \{0,1\}$ and $x \in \mathcal{X}$,*

$$\mathbb{P}\{Y^*(t) = 1 | T^* = 1, X^* = x\} = \mathbb{P}\{Y^*(t) = 1 | T^* = 0, X^* = x\}.$$

The first requirement of assumption B implies that the potential outcome $Y^*(t)$ cannot be 0 or 1 with probability 1 for some value of x. The second condition of assumption B is the standard overlap condition in the literature. Assumption C says that the potential outcomes $Y^*(1)$ and $Y^*(0)$ are conditionally independent of the treatment T^* given $X^* = x$.

We now provide the following identification result in the spirit of Holland and Rubin (1988).

Theorem 1 (Holland and Rubin (1988))**.** *Suppose that assumptions B and C are satisfied. Then, under design 1, we have $\theta_{\text{OR}}(x) = \text{OR}^*(x) = \text{OR}(x)$ for all $x \in \mathcal{X}$; under design 2, we have $\theta_{\text{RR}}(x) = \text{RR}^*(x) = \text{OR}(x)$ for all $x \in \mathcal{X}$.*

Theorem 1 slightly extends the result of Holland and Rubin (1988); they did not consider design 2, but their arguments can be used in a straightforward manner. In substance, the observed odds ratio $\text{OR}(x)$ identifies the causal odds ratio $\theta_{\text{OR}}(x)$ under design 1 and the causal relative risk $\theta_{\text{RR}}(x)$ under design 2. One practical message of theorem 1 is that it might be more beneficial to sample a control group from the unconditional population if a researcher cares mainly about $\theta_{\text{RR}}(x)$. In light of this, we may regard designs 1 and 2 as studies suitable for the causal odds ratio and causal relative risk, respectively.

2.4. **Causal Interpretation without Strong Ignorability.** Strong ignorability is convenient but it may be too strong for observational data; T^* is often a deliberate decision of an individual agent. In this subsection, we establish an alternative causal interpretation of OR using the framework of partial identification. In particular, we build on assumptions of monotone treatment response (Manski, 1997) and monotone treatment selection (Manski and Pepper, 2000).

Assumption D (Monotone Treatment Response)**.** *We have* $Y^*(1) \geq Y^*(0)$ *almost surely.*

Assumption E (Monotone Treatment Selection)**.** *For all* $t \in \{0, 1\}$ *and* $x \in \mathcal{X}$*,*

$$\mathbb{P}\{Y^*(t) = 1 | T^* = 1, X^* = x\} \geq \mathbb{P}\{Y^*(t) = 1 | T^* = 0, X^* = x\}.$$

Assumption D rules out the possibility of $Y^*(1) = 0$ and $Y^*(0) = 1$ almost surely. For instance, if an individual, who is randomly assigned to higher education, does not earn high incomes, then he or she will not be highly paid, either, when randomly assigned to no higher education. Assumption E says that other things being equal, those who have higher degrees are at least as likely to earn high incomes, if their education attainment was randomly assigned, compared to those who did not have higher degrees.

The following theorem gives the partial identification results for $\theta_{\text{RR}}(x)$ and $\theta_{\text{OR}}(x)$.

Theorem 2. *Suppose that assumption B holds. The following inequalities are sharp.*

(i) If assumption D is satisfied, then $1 \leq \theta_{\text{RR}}(x) \leq \theta_{\text{OR}}(x)$ *for all* $x \in \mathcal{X}$ *under each of the two sampling designs.*

(ii) If assumption E is satisfied, then $\theta_{\text{OR}}(x) \leq \text{OR}(x)$ *under design 1 and* $\theta_{\text{RR}}(x) \leq \text{OR}(x)$ *under design 2.*

(iii) The following two statements are equivalent:

(a) $\theta_{\text{OR}}(x) = \text{OR}(x)$ *in design 1 and* $\theta_{\text{RR}}(x) = \text{OR}(x)$ *in design 2;*

(b) Assumption E is satisfied with equality, i.e. assumption C holds.

Parts (i) and (ii) of Theorem 2 imply that if assumptions D and E are satisfied, then $\text{OR}(x)$ can be understood as a sharp upper bound of causal relative risk both under designs 1 and 2. More specifically, for all $x \in \mathcal{X}$, we have

$$1 \leq \theta_{\text{RR}}(x) \leq \theta_{\text{OR}}(x) \leq \text{OR}(x) \qquad \text{under design 1;} \tag{7}$$

$$1 \leq \theta_{\text{RR}}(x) \leq \text{OR}(x) \qquad \text{under design 2.} \tag{8}$$

Theorems 1 and 2 articulate how to give causal interpretation to $\mathrm{OR}(x)$ in general. Assumption E allows for assumption C as a special case. Indeed, theorem 2 shows that point identification holds if and only if the unconfoundedness condition is satisfied.

Assumptions D and E are not individually testable, but they jointly have a testable implication, i.e. $\mathrm{OR}(x) \geq 1$ for all $x \in \mathcal{X}$ by theorem 2, for which a nonparametric test can be constructed via the general framework of testing functional inequalities (see, e.g., Chernozhukov, Lee, and Rosen, 2013; Lee, Song, and Whang, 2018).

In case-control studies, it is commonly assumed that there is some $\epsilon > 0$ such that $0 < \mathbb{P}(Y^* = 1 | X^* = x) \leq \epsilon$ for all $x \in \mathcal{X}$. When we consider the case where $\epsilon \to 0$, we refer to this condition as the rare-disease assumption (e.g. Breslow, 1996; Manski, 2009). The rare-disease assumption leads to $|\theta_{\mathrm{RR}}(x) - \theta_{\mathrm{OR}}(x)| \to 0$ as $\epsilon \to 0$. Hence, if both strong ignorability and rare-disease are assumed, then $\theta_{\mathrm{RR}}(x)$ is well-approximated by $\mathrm{OR}(x)$ under design 1.

However, our identification analysis shows that a researcher does not have to resort to strong ignorability, nor to the the rare-disease assumption, in order to provide $\mathrm{OR}(x)$ with causal interpretation. If both MTR and MTS conditions are plausible, then a researcher can interpret $\mathrm{OR}(x)$ as the (sharp) upper bound of the causal relative risk $\theta_{\mathrm{RR}}(x)$ under both designs 1 and 2.

2.5. **Heterogeneity and Aggregation.** The functional parameter $\mathrm{OR}(x)$ is difficult to estimate nonparametrically with good precision when the dimension of X is high. To avoid the curse of dimensionality, it is popular in case-control studies to adopt logistic regression at the true population level: that is,

$$\mathbb{P}(Y^* = 1 | T^* = t, X^* = x) = \frac{\exp(\alpha_0 + t\alpha_1 + x^\top \alpha_2 + t x^\top \alpha_3)}{1 + \exp(\alpha_0 + t\alpha_1 + x^\top \alpha_2 + t x^\top \alpha_3)}, \tag{9}$$

which implies that

$$\alpha_1 + x^\top \alpha_3 = \log\{\mathrm{OR}^*(x)\} = \log\{\mathrm{OR}(x)\}$$

for all $x \in \mathcal{X}$; therefore, under the rare-disease assumption, $\log\{\mathrm{RR}^*(x)\} \approx \alpha_1 + x^\top \alpha_3$ as well.

The parametric assumption is popular, but it is restrictive. For instance, the formulation in equation (9) limits the possible forms of heterogeneous causal effects; without the parametric assumption, $\log\{\mathrm{OR}(x)\}$ is generally an unknown function of x that can be highly nonlinear. In this paper we take a nonparametric approach, where we aim at estimating $\mathrm{OR}(x)$ without using any parametric assumption, after which we aggregate it by integrating over x.

Since $F_{X|Y}(\cdot|1)$ and $F_{X|Y}(\cdot|0)$ are identified in our study designs, we consider

$$\beta(y) := \int_{\mathcal{X}} \log\{\mathrm{OR}(x)\} dF_{X|Y}(x|y) \text{ for } y = 0, 1, \tag{10}$$

which is the weighted average of the log odds ratio using $F_{X|Y}(x|y)$ as weights: the argument y indicates which distribution of X, and hence which distribution of X^*, is used to aggregate the log odds ratio. Specifically, under design 2, $\beta(0)$ is equal to $\mathbb{E}\left[\log\{\mathrm{OR}(X^*)\}\right]$. Under design 1, if the population fraction of the case (i.e. $\mathbb{P}(Y^* = 1)$) is known to the researcher, which has been frequently assumed in the econometrics literature (since Manski and Lerman, 1977), then $\mathbb{E}\left[\log\{\mathrm{OR}(X^*)\}\right]$ can be obtained by taking the weighted average of $\beta(y)$, i.e.

$$\mathbb{E}\left[\log\{\mathrm{OR}(X^*)\}\right] = \beta(1)\mathbb{P}(Y^* = 1) + \beta(0)\mathbb{P}(Y^* = 0).$$

If $\mathbb{P}(Y^* = 1)$ is unknown but only its upper bound is known, then we can undertake a bound analysis on $\mathbb{E}\left[\log\{\mathrm{OR}(X^*)\}\right]$ by using $\beta(1)$ and $\beta(0)$; this problem will be further discussed in section 5. Therefore, in the next two sections, we will treat $\beta(y)$ as the main estimand of interest; our discussion on semiparametric efficiency and machine-learning approaches will focus on $\beta(y)$. It relies on the researcher's view on assumptions C to E whether it is the aggregation of the logarithm of the causal parameter itself or its sharp identifiable upper bound. Using our proposed estimators, we discuss how to carry out causal inferences in section 5.

In equation (10), the logarithm is taken before aggregating the odds ratio; alternatively, one may take an expectation of the odds ratio directly. This case can be handled similarly. See Appendix A for details.

3. Efficient Influence Function for $\beta(y)$

We consider estimating the parameter $\beta(y)$, for which we do not impose any parametric restrictions anywhere. As a first step, we derive the semiparametric efficiency bound under both designs 1 and 2; since the mathematical structure of the likelihood function is the same, we do not need to distinguish design 1 from design 2. For this purpose, we will use the generic notation using the observed variables (Y, T, X) instead of the original random variables of interest, i.e. (Y^*, T^*, X^*). We start with the following assumptions for regularity.

Assumption F (Bounded Probabilities)**.** *There is a constant $\varepsilon > 0$ such that for each $y = 0, 1$, $\varepsilon \leq \mathbb{P}(T = 1|X, Y = y) \leq 1 - \varepsilon$ and $\varepsilon \leq \mathbb{P}(Y = 1|X) \leq 1 - \varepsilon$ almost surely.*

Assumption G (Regular Distribution)**.** *The distribution function $F_{X|Y}$ has a probability density $f_{X|Y}$ that satisfies $0 < f_{X|Y}(x|y) < \infty$ for all $x \in \mathcal{X}$ and $y = 0, 1$.*

Assumptions F and G are, in principle, testable since they are about the random variables observed in the sample. Assumption F is slightly stronger than what we need to derive the efficient influence function, but it will be needed to establish statistical properties of our proposed estimators later. Assumption G focuses on the case where X is continuous but this is only for the sake of notational simplicity; if X is discrete or mixed, then $f_{X|Y}$ should be understood as a general Radon-Nikodym density with respect to some dominating measure.

Under the Bernoulli sampling scheme, the likelihood of a single observation (Y, T, X) is given by

$$L(Y, T, X) = \big\{(1 - h_0)\mathcal{P}_0(T, X)\big\}^{1-Y}\big\{h_0\mathcal{P}_1(T, X)\big\}^{Y}, \tag{11}$$

where for $y = 0, 1$,

$$\mathcal{P}_y(T, X) = f_{X|Y}(X|y)\mathbb{P}(T = 1|X, Y = y)^T \big\{1 - \mathbb{P}(T = 1|X, Y = y)\big\}^{1-T}. \quad (12)$$

The likelihood in equation (11) is a simple mixture of two binary likelihoods. The tangent space can be derived by using regular parametric submodels $\mathcal{P}_y(T, X; \gamma)$ such that $\mathcal{P}_y(T, X; \gamma_0) = \mathcal{P}_y(T, X)$ for $y = 0, 1$. The tangent space is described in the following lemma.

Lemma 2. *Consider the Bernoulli sampling scheme of design 1 or design 2. The tangent space is given by the set of functions of the following form:*

$$s(Y, T, X) = (1 - Y)\Big[a_0(X) + \big\{T - \mathbb{P}(T = 1|X, Y = 0)\big\}b_0(X)\Big] \\ + Y\Big[a_1(X) + \big\{T - \mathbb{P}(T = 1|X, Y = 1)\big\}b_1(X)\Big],$$

where the functions a_y and b_y are such that $\mathbb{E}\{a_y(X)|Y = y\} = 0$ and $\mathbb{E}\{s^2(Y, T, X)\} < \infty$ for each $y = 0, 1$.

The following theorem shows that $\beta(y)$ is pathwise differentiable along the regular parametric submodels at γ_0 in the sense of Newey (1990, 1994). Before we present the theorem, define

$$w(X) := \frac{f_{X|Y}(X|0)}{f_{X|Y}(X|1)}. \quad (13)$$

Further, for $y = 0, 1$, define

$$\Delta_y(Y, T, X) := \frac{Y^y(1 - Y)^{1-y}\{T - \mathbb{P}(T = 1|X, Y = y)\}}{\mathbb{P}(T = 1|X, Y = y)\{1 - \mathbb{P}(T = 1|X, Y = y)\}}.$$

We establish the following result using the approach taken by Hahn (1998).

Theorem 3. *Suppose that assumptions A, F and G hold and that we have a sample by Bernoulli sampling. Then, for $y = 0, 1$, $\beta(y)$ is pathwise differentiable and its pathwise* *derivative is given by*

$$F_y(Y,T,X) = \frac{Y^y(1-Y)^{1-y}}{h_0^y(1-h_0)^{1-y}}\Big\{\log \mathrm{OR}(X) - \beta(y)\Big\} - \frac{\Delta_0(Y,T,X)}{(1-h_0)w(X)^y} + \frac{w(X)^{1-y}\Delta_1(Y,T,X)}{h_0}.$$

Further, F_y is an element of the tangent space, and therefore, the semiparametric efficiency bound for $\beta(y)$ is given by $\mathbb{E}\{F_y^2(Y,T,X)\}$.

Theorem 3 shows the efficiency bound for $\beta(y)$, and it also implies that the asymptotic variance of a $\sqrt{n}$–consistent and asymptotically linear estimator of $\beta(y)$ should be $\mathbb{E}\{F_y^2(Y,T,X)\}$ by Theorem 2.1 of Newey (1994). Since $\beta(y)$ is the expectation of $\log \mathrm{OR}(X)$ with respect to the distribution of X given $Y = y$, it satisfies

$$\mathbb{E}\Big\{\log \mathrm{OR}(X) - \beta(y)\big|Y = y\Big\} = \mathbb{E}\left[\frac{Y^y(1-Y)^{1-y}}{h_0^y(1-h_0)^{1-y}}\Big\{\log \mathrm{OR}(X) - \beta(y)\Big\}\right] = 0, \quad (14)$$

which is the expected value of the first term that appears in $F_y(Y,T,X)$; the other terms in $F_y(Y,T,X)$ are for adjustment to address the effect of first step nonparametric estimation of $\log \mathrm{OR}(X)$ via $\mathbb{P}(T = 1|X = x, Y = y)$.

4. Efficient Estimation of $\beta(y)$

Efficient estimators of $\beta(y)$ for $y = 0, 1$ can be constructed in multiple ways. The most straightforward approach is just using equation (14), i.e. we base an estimator on

$$\beta(y) = \mathbb{E}\left[\frac{Y^y(1-Y)^{1-y}}{h_0^y(1-h_0)^{1-y}} \log \mathrm{OR}(X)\right], \quad (15)$$

where we plug in a nonparametric estimator of $\mathrm{OR}(x)$. Alternatively, we may include the adjustment terms upfront to use $\mathbb{E}\{F_y(Y,T,X)\} = 0$ as an estimating equation. In other words, we may estimate $\beta(y)$ by constructing a sample analog estimator from the following alternative expression: $\beta(y)$ is equal to

$$\mathbb{E}\left[\frac{Y^y(1-Y)^{1-y}}{h_0^y(1-h_0)^{1-y}}\log \mathrm{OR}(X) - \frac{\Delta_0(Y,T,X)}{(1-h_0)w(X)^y} + \frac{w(X)^{1-y}\Delta_1(Y,T,X)}{h_0}\right]. \tag{16}$$

This approach requires additional (nonparametric) estimation of $w(X)$, but since $\mathbb{E}\{\Delta_y(Y,T,X)|X\} = 0$ almost surely for $y = 0, 1$, having an incorrect function for $w(X)$ does not matter for the consistency of the estimator based on equation (16).[3]

Suppose that we have the sample $\{(Y_i, T_i, X_i) : i = 1, \ldots, n\}$, where (Y_i, T_i, X_i)'s are i.i.d. copies of (Y, T, X). Using this sample, we propose sieve logistic estimators based on equation (15) in section 4.1. In section 4.2, we show that the moment condition in equation (16) satisfies Neyman orthogonality in the sense of Chernozhukov, Chetverikov, Dimirer, Duflo, Hansen, Newey, and Robins (2018, DML hereafter). This leads to double/debiased machine learning (DML) estimators, which we present in section 4.3. Throughout the discussion we assume that h_0 is known since it is part of the sampling scheme. However, if it is unknown, then using $\hat{h} = \sum_{i=1}^n Y_i/n$ instead of h_0 does not change the first-order asymptotic behaviors of the estimators based on (15) and (16), as long as $\mathcal{P}_0$ and $\mathcal{P}_1$ do not depend on h_0.

4.1. **Retrospective Sieve Logistic Estimation.** Recall that the observed odds ratio in equation (4) can be expressed as

$$\mathrm{OR}(x) = \frac{\mathbb{P}(T=1|X=x,Y=1)}{\mathbb{P}(T=0|X=x,Y=1)}\frac{\mathbb{P}(T=0|X=x,Y=0)}{\mathbb{P}(T=1|X=x,Y=0)}.$$

We model the treatment probabilities by infinite dimensional logistic regression: i.e. for $y = 0, 1$,

$$\mathbb{P}(T=1|X=x,Y=y) = \frac{\exp\left(\sum_{j=1}^{\infty}\phi_j(x)\mu_{j,y}\right)}{1+\exp\left(\sum_{j=1}^{\infty}\phi_j(x)\mu_{j,y}\right)},$$

[3]Misspecification of $w(X)$ may affect the asymptotic distribution of our proposed estimator. We limit our attention to nonparametric estimation of $w(X)$ to minimize the possibility of misspecification.

where $\{\phi_j : j = 1, 2, \ldots\}$ is a series of basis functions and $\{\mu_{j,y} : j = 1, 2, \ldots\}$ is a series of unknown coefficients for each $y = 0, 1$. It then follows that for each $y = 0, 1$,

$$\log \frac{\mathbb{P}(T = 1|X = x, Y = y)}{\mathbb{P}(T = 0|X = x, Y = y)} = \sum_{j=1}^{\infty} \phi_j(x)\mu_{j,y}. \tag{17}$$

Therefore, by using equation (15) and assumption F, we obtain

$$\begin{aligned} \beta(y) &= \sum_{j=1}^{\infty} \int_{\mathcal{X}} \phi_j(x) dF_{X|Y}(x|y) \left(\mu_{j,1} - \mu_{j,0}\right) \\ &\approx \sum_{j=1}^{J_n} \int_{\mathcal{X}} \phi_j(x) dF_{X|Y}(x|y) \left(\mu_{j,1} - \mu_{j,0}\right), \end{aligned} \tag{18}$$

provided that J_n diverges to infinity as $n \to \infty$. Equation (18) suggests the following two-step sieve estimation strategy:

(i) In the first step, for each $y = 0, 1$, estimate $\{\mu_{j,y} : y = 0, 1\ j = 1, \ldots, J_n\}$ by logistic regression of T_i on $\{\phi_j(X_i) : j = 1, \ldots, J_n\}$ with the $Y_i = y$ sample.

(ii) In the second step, construct a sample analog of equation (18): i.e.

$$\widehat{\beta}(y) := \sum_{j=1}^{J_n} \int_{\mathcal{X}} \phi_j(x) d\widehat{F}_{X|Y}(x|y) \left(\widehat{\mu}_{j,1} - \widehat{\mu}_{j,0}\right), \tag{19}$$

where $\widehat{\mu}_{j,y}$'s are sieve logit estimates from the first step and

$$\int_{\mathcal{X}} \phi_j(x) d\widehat{F}_{X|Y}(x|y) = \frac{\sum_{i=1}^{n} Y_i^d (1 - Y_i)^d \phi_j(X_i)}{\sum_{i=1}^{n} Y_i^d (1 - Y_i)^d}.$$

Since the retrospective probability model is used in equation (17), we call the estimator defined in (19) the *retrospective sieve logistic estimator* of $\beta(y), y = 0, 1$. It can be computed using standard software for logistic regression, as described in algorithm 1.

The procedure described in algorithm 1 achieves the first step by running a combined logistic regression of T_i on Y_i, the sieve basis terms and the interactions between Y_i and the sieve basis terms. This is first-order equivalent since Y_i is binary

Algorithm 1: Retrospective Sieve Logistic Estimator of $\beta(1)$

Input: $\{(Y_i, T_i, X_i) : i = 1, \ldots, n\}$, tuning parameter J_n and basis functions $\{\phi_j(\cdot) : j = 1, \ldots, J_n\}$

Output: estimate of $\beta(1)$ and its standard error

1. Construct $\{\phi_1(X_i), \ldots, \phi_{J_n}(X_i) : i = 1, \ldots, n\}$, where an intercept term is excluded in ϕ_j's;
2. For each $j = 1, \ldots, J_n$, compute the empirical mean of $\phi_j(X_i)$ using only the case sample ($Y_i = 1$) and construct the demeaned version, say $\varphi_j(X_i)$, of $\phi_j(X_i)$;
3. Run a logistic regression of T_i on the following regressors: an intercept term, Y_i, $\varphi_j(X_i), j = 1, \ldots, J_n$, and interactions between Y_i and $\varphi_j(X_i), j = 1, \ldots, J_n$, using standard software;
4. Read off the estimated coefficient for Y_i and its standard error

and full interaction terms are included. For the second step, instead of evaluating the right-hand side of equation (19) after logistic regression, $\phi_j(X_i)$'s are demeaned first using only the case sample so that the resulting coefficient for Y_i is first-order equivalent to the estimator defined in equation (19). The advantage of the formulation in algorithm 1 is that the standard error of $\widehat{\beta}(1)$ can be read off directly from standard software without any further programming. It is straightforward to modify algorithm 1 for estimating $\beta(0)$. One has to compute the empirical mean of $\phi_j(X_i)$ using only the control sample ($Y_i = 0$) for the demeaning step.

Sieve logistic estimators have been popular in the literature, including the propensity score estimator used in Hirano, Imbens, and Ridder (2003). To the best of our knowledge, it is novel to adopt retrospective sieve logistic estimators in the context of case-control studies. It is not difficult to work out formal asymptotic properties of our proposed sieve estimator in view of the well-established literature on two-step sieve estimation (see, e.g., Ai and Chen, 2003, 2012; Ackerberg, Chen, Hahn, and Liao, 2014, among many others). Furthermore, conventional normal inference based on the standard error obtained in algorithm 1 is valid for semiparametric inference (e.g. Ackerberg, Chen, and Hahn, 2012). For brevity of the paper, we omit details.

4.2. **Neyman Orthogonality.** Both of the estimating equations in (15) and (16) depend on nonparametric objects that need to be estimated in advance. Equation (15) is simpler but equation (16) has an advantage that it is robust to local perturbation on the unknown functions that are estimated in the first step. It requires extra notation to discuss this result formally.

Let $\mathcal{W}$ be the set of functions on $\mathcal{X}$ that are bounded and bounded away from zero. Similarly, let $\mathcal{G}$ be the set of functions $g : \mathcal{X} \to [\epsilon,\ 1-\epsilon]$ for some $\epsilon > 0$. For $\eta = (\eta_1^\intercal, \eta_2)^\intercal$ with $\eta_1 = (a, b)^\intercal \in \mathcal{G}^2$ and $\eta_2 \in \mathcal{W}$, define

$$\widetilde{\mathrm{OR}}(\eta_1)[X] = \big[b(X)\{1-a(X)\}\big] / \big[\{1-b(X)\}a(X)\big] \quad \text{and} \quad \tilde{w}(\eta_2)[X] = \eta_2(X).$$

So, $\widetilde{\mathrm{OR}}(\cdot)[X]$ and $\tilde{w}(\cdot)[X]$ denote (candidate) mappings from $\mathcal{G}^2$ and $\mathcal{W}$, respectively, such that they are equal to $\mathrm{OR}(X)$ and $w(X)$ when they are evaluated at $\eta_{10} \in \mathcal{G}^2$ and $\eta_{20} \in \mathcal{W}$, respectively, where $\eta_{10}(x) = \big(\mathbb{P}(T = 1|X = x, Y = 0), \mathbb{P}(T = 1|X = x, Y = 1)\big)^\intercal$ and $\eta_{20}(x) = w(x)$. Now, we define the mapping $\tilde{F}_y(\cdot)[Y, T, X]$ by

$$\begin{aligned}\tilde{F}_y(\eta)[Y,T,X] &:= \frac{Y^y(1-Y)^{1-y}}{h_0^y(1-h_0)^{1-y}} \log \widetilde{\mathrm{OR}}(\eta_1)[X] - \beta(y) \\ &+ \frac{Y}{h_0}\{\tilde{w}(\eta_2)[X]\}^{1-y} \frac{\{T-b(X)\}}{b(X)\{1-b(X)\}} - \frac{1-Y}{1-h_0}\{\tilde{w}(\eta_2)[X]\}^{-y} \frac{\{T-a(X)\}}{a(X)\{1-a(X)\}},\end{aligned}$$

where $\eta = (a, b, \eta_2)^\intercal \in \mathcal{G}^2 \times \mathcal{W}$. So, we have $\tilde{F}_y(\eta_0)[Y, T, X] = F_y(Y, T, X)$, where $\eta_0 = (\eta_{10}^\intercal, \eta_{20})^\intercal$. We are now ready to state the main theorem of this subsection.

Theorem 4. *Suppose that assumptions F and G hold. Then, under both designs 1 and 2, and for each $y = 0, 1$, the Gateaux derivative of $\tilde{F}_y(\cdot)[Y, T, X]$ at η_0 has mean zero: i.e.*

$$\mathbb{E}\Big[\partial_\gamma \tilde{F}_y\{\eta_0 + \gamma(\eta - \eta_0)\}[Y,T,X]\Big|_{\gamma=0}\Big] = 0$$

for all $\eta \in \mathcal{G}^2 \times \mathcal{W}$.

Theorem 4 says that $F_y(Y,T,X)$ provides a Neyman orthogonal moment function. The fact that small perturbations around η_0 do not have first-order asymptotic consequences is known as the local robustness property. In this case, the first step nonparametric estimation does not have any first-order effect, i.e. the limiting distribution would be the same as if η_0 were known, because all the adjustment terms that are needed to address the effect of the first step estimation are already reflected in F_y.

4.3. **Retrospective Double/Debiased Machine Learning Estimation.** When the dimension of X is higher than the sample size, it is infeasible to implement the sieve estimator proposed in section 4.1. In this section we consider using machine-learning-based estimators in the first step, which will allow X to be of high dimension. In view of Neyman orthognality established in section 4.2, we build a new estimator based on equation (16), which requires estimation of $w(x)$ defined in equation (13). In high-dimensional settings, it would be impractical to estimate $f_{X|Y}(x|0)$ and $f_{X|Y}(x|1)$ separately and to take the ratio to obtain an estimator of $w(x)$. Instead, we use the Bayes rule to obtain[4]

$$w(x) = \frac{f_{X|Y}(x|0)}{f_{X|Y}(x|1)} = \frac{\mathbb{P}(Y=0|X=x)}{\mathbb{P}(Y=1|X=x)} \frac{h_0}{1-h_0}, \tag{20}$$

which suggests that we estimate $\mathbb{P}(Y=1|X=x)$ since $h_0 = \mathbb{P}(Y=1)$ is either known or trivial to estimate. The key insight here is that it may be unrealistic to assume the sparsity of $f_{X|Y}(x|y)$ for each $y=0,1$, but $w(x)$ can be estimated by sparsity-based models since the sparsity of $w(x)$ is equivalent to that of $\mathbb{P}(Y=0|X=x)/\mathbb{P}(Y=1|X=x)$. Therefore, we may rely on machine-learning methods to estimate not only $\mathbb{P}(T=1|X=x,Y=y), y=0,1$, but also $\mathbb{P}(Y=1|X=x)$. For example, we may use ℓ_1-penalized logistic estimation for estimating all relevant probability models to construct an estimator of $\beta(y), y=0,1$. Specifically, in order

[4] Heckman and Todd (2009) use the same relationship in the context of propensity score matching under treatment-based sampling.

to develop an efficient estimator that works in the high-dimensional settings, we build on DML.

Let $K \geq 2$ be some fixed integer (say, 5, 10 or 20). For simplicity, assume that n is divisible by K. Let $\{I_k : k = 1, \dots, K\}$ denote a K-fold partition of $\{1, \dots, n\}$ such that $|I_k| = n/K$ for each k. Suppose that one estimates $\eta_0 = (\eta_{10}^\intercal, \eta_{20}^\intercal)^\intercal$ using a machine-learning estimator, say $\hat{\eta}_k$, using observations that belong to $I_k^c := \{1, \dots, n\} \setminus I_k$ for each k. Then, the retrospective double/debiased machine learning estimator $\widehat{\beta}_{\mathrm{DML}}(y)$ of $\beta(y)$, $y = 0, 1$, is defined by

$$\widehat{\beta}_{\mathrm{DML}}(y) := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} \widehat{\psi}_{i,k}(y), \tag{21}$$

where $\hat{h} := n^{-1} \sum_{i=1}^{n} Y_i$,

$$\begin{aligned} \widehat{\psi}_{i,k}(y) := \frac{Y_i^y (1 - Y_i)^{1-y}}{\hat{h}^y (1 - \hat{h})^{1-y}} \log \widetilde{\mathrm{OR}}(\hat{\eta}_{1,k})[X_i] + \frac{Y_i}{\hat{h}} \{\tilde{w}(\hat{\eta}_{2,k})[X_i]\}^{1-y} \frac{\{T_i - \hat{p}_{1,k}(X_i)\}}{\hat{p}_{1,k}(X_i)\{1 - \hat{p}_{1,k}(X_i)\}} \\ - \frac{(1 - Y_i)}{(1 - \hat{h})} \{\tilde{w}(\hat{\eta}_{2,k})[X_i]\}^{-y} \frac{T_i - \hat{p}_{0,k}(X_i)}{\hat{p}_{0,k}(X_i)\{1 - \hat{p}_{0,k}(X_i)\}}, \end{aligned} \tag{22}$$

and

$$\begin{aligned} \hat{p}_{1,k}(x) &:= \widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 1 | X = x, Y = 1), \\ \hat{p}_{0,k}(x) &:= \widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 1 | X = x, Y = 0), \\ \widetilde{\mathrm{OR}}(\hat{\eta}_{1,k})[x] &:= \frac{\widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 1 | X = x, Y = 1)}{\widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 0 | X = x, Y = 1)} \frac{\widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 0 | X = x, Y = 0)}{\widehat{\mathbb{P}}_{\mathrm{ML},k}(T = 1 | X = x, Y = 0)}, \\ \tilde{w}(\hat{\eta}_{2,k})[x] &:= \frac{\widehat{\mathbb{P}}_{\mathrm{ML},k}(Y = 0 | X = x)}{\widehat{\mathbb{P}}_{\mathrm{ML},k}(Y = 1 | X = x)} \frac{\hat{h}}{(1 - \hat{h})}. \end{aligned}$$

Here, $\widehat{\mathbb{P}}_{\mathrm{ML},k}$ denotes a machine-learning estimator of a probability model using observations that belong to I_k^c. We summarize the estimation procedure in algorithm 2.

Algorithm 2: Retrospective Double/Debiased Machine Learning Estimator of $\beta(y), y = 0, 1$

Input: $\{(Y_i, T_i, X_i) : i = 1, \ldots, n\}$, K, machine learning methods for estimating probability models

Output: estimate of $\beta(1)$ and its standard error

1 Construct a K-fold partition $\{I_k : k = 1, \ldots, K\}$ of $\{1, \ldots, n\}$ of approximately equal size;

2 For each k, use observations belonging to I_k^c to obtain machine learning estimates $\hat{\eta}_k$ of $\mathbb{P}(T = 1|X = x, Y = 1)$, $\mathbb{P}(T = 1|X = x, Y = 0)$ and $\mathbb{P}(Y = 1|X = x)$, respectively;

3 For each k, use observations belonging to I_k to construct $\widehat{\psi}_{i,k}(y)$ in equation (22);

4 Obtain the estimate of $\beta(1)$ by equation (21) and its standard error $\widehat{\sigma}_{\text{DML}}(y)/\sqrt{n}$ by

$$\widehat{\sigma}^2_{\text{DML}}(y) := \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i \in I_k}\left\{\widehat{\psi}_{i,k}(y) - \widehat{\beta}_{\text{DML}}(y)\right\}^2. \tag{23}$$

Let $\|\cdot\|_{P,2}$ denote the $L_2(P)$-norm, where P is a probability distribution that (Y, T, X) takes: i.e.

$$\|a\|^2_{P,2} = \max_{1\le \ell \le d}\left\{\mathbb{E}[a_\ell^2(Y, T, X)]\right\}^{1/2}$$

for a d-dimensional vector-valued function $a := (a_1, \ldots, a_d)$.

Assumption H (First-Stage Estimation)**.** *There exist sequences $\delta_n \ge n^{-1/2}$ and τ_n of positive constants both approaching zero such that for each $k = 1, \ldots, K$, $\|\hat{\eta}_k - \eta_0\|_{P,2} \le \delta_n n^{-1/4}$ with probability no less than $1 - \tau_n$.*

Assumption H resembles classical rate requirements in semiparametric estimation. General theory of DML allows for a general norm; however, the $L_2(P)$-norm is the most convenient for machine learning estimators. The required rate is attainable for a variety of machine learning methods. For instance, the primitive conditions for ℓ_1-penalized logit estimators are worked out by van de Geer (2008) and Belloni, Chernozhukov, and Wei (2016) among others.

An application of Theorems 3.1 and 3.2 of DML gives the following result that formally justifies the estimation method proposed in algorithm 2.

Theorem 5. *Let $\{\mathcal{P}_n : n \geq 1\}$ be a sequence of sets of probability distributions of (Y, T, X). Suppose that for all $n \geq 3$ and $P \in \mathcal{P}_n$, (16) and assumptions F to H hold and that we have a sample by the Bernoulli sampling scheme of design 1 or design 2. Then, for $y = 0, 1$,*

$$\sqrt{n}\frac{\{\widehat{\beta}_{\mathrm{DML}}(y) - \beta(y)\}}{\widehat{\sigma}_{\mathrm{DML}}(y)} \to_d \mathbb{N}(0,1) \quad \textit{uniformly over } P \in \mathcal{P}_n,$$

and $\widehat{\sigma}^2_{\mathrm{DML}}(y) \to_p \mathbb{E}\left[F_y^2(Y,T,X)\right]$ uniformly over $P \in \mathcal{P}_n$, where $\widehat{\sigma}^2_{\mathrm{DML}}(y)$ is defined in equation (23).

5. Discussion: The Main Takeaway and Inferential Issues

In this section we discuss and summarize some of the important messages from our findings. Recall that the main estimand of interest is $\beta(y)$ for $y = 0, 1$, which is an aggregated version of $\log\{\mathrm{OR}(x)\}$. With causal inference in mind, the corresponding causal parameters would be either

$$\xi_{\mathrm{RR}}(y) := \int_{\mathcal{X}} \log\{\theta_{\mathrm{RR}}(x)\} dF_{X|Y}(x|y) \quad \text{or} \quad \xi_{\mathrm{OR}}(y) := \int_{\mathcal{X}} \log\{\theta_{\mathrm{OR}}(x)\} dF_{X|Y}(x|y).$$

Here, we note that $\log\theta_{\mathrm{RR}}(x)$ is easier to interpret than $\log\theta_{\mathrm{OR}}(x)$, so the former is a more natural causal parameter to target. Also, it is arguably more desirable to aggregate $\log\theta_{\mathrm{RR}}(x)$ by the true distribution of X^*: i.e.

$$\bar{\xi}_{\mathrm{RR}} := \mathbb{E}\big[\log\mathbb{P}\{Y^*(1) = 1|X^*\}\big] - \mathbb{E}\big[\log\mathbb{P}\{Y^*(0) = 1|X^*\}\big], \tag{24}$$

and we have

$$\bar{\xi}_{\mathrm{RR}} = \begin{cases} \xi_{\mathrm{RR}}(0)(1-p^*) + \xi_{\mathrm{RR}}(1)p^* & \text{under design 1,} \\ \xi_{\mathrm{RR}}(0) & \text{under design 2,} \end{cases}$$

where $p^* := \mathbb{P}(Y^* = 1)$.

In this setup, causal inference can be understood as how we relate the estimand $\beta(y)$ with $\xi_{\mathrm{RR}}(y)$, and eventually with $\bar{\xi}_{\mathrm{RR}}$, for which we need to address the fact that p^* is unidentified. Below we discuss each step in detail.

How to relate $\beta(y)$ with $\zeta(y)$ depends on several assumptions as well as the sampling design itself. In the case of case-control sampling (i.e. design 1), strong ignorability ensures that $\beta(y) = \zeta_{\mathrm{OR}}(y)$ but we do not learn about $\zeta_{\mathrm{RR}}(y)$ from $\beta(y)$, unless the rare-disease assumption is additionally in place: if the case is rare in the population (uniformly across the values of X^*), then $\zeta_{\mathrm{OR}}(y)$ is a good approximation of $\zeta_{\mathrm{RR}}(y)$. The case of case-population sampling (i.e. design 2) is easier, because strong ignorability is sufficient to guarantee $\beta(0) = \zeta_{\mathrm{RR}}(0)$. Therefore, a confidence interval for $\zeta_{\mathrm{RR}}(0)$ in this case can be computed in the usual symmetric and two-sided way by using any of the proposed estimators of $\beta(0)$.

If strong ignorability is not credible, then the (approximate) equality relationship between $\beta(y)$ and $\zeta_{\mathrm{RR}}(y)$ breaks down. However, we have shown that if the MTR and MTS conditions are satisfied, we have

$$0 \leq \zeta_{\mathrm{RR}}(y) \leq \beta(y) \tag{25}$$

under both designs 1 and 2, where the inequalities are sharp. Further, these inequalities do not require the rare disease assumption, and hence they are robust against its violation. Equation (25) implies that an estimate of $\beta(y)$, e.g. $\widehat{\beta}_{\mathrm{DML}}(y)$, should be interpreted carefully: a large estimate does not necessarily confirm a large causal effect but a small estimate does confirm a small causal effect. Also, a confidence interval for $\zeta_{\mathrm{RR}}(y)$ should be computed differently. For example, if $\widehat{\beta}_{\mathrm{DML}}(y)$ is used, then an asymptotically valid confidence interval for $\zeta_{\mathrm{RR}}(y)$ should be computed by $\left[0,\ \widehat{\beta}_{\mathrm{DML}}(y) + z_{1-\alpha} \cdot \widehat{\sigma}_{\mathrm{DML}}(y)\right]$, where $z_{1-\alpha}$ is a one-sided standard normal critical value. Since $\widehat{\beta}_{\mathrm{DML}}(y)$ is an efficient estimator, it will lead to a tight one-sided confidence interval for $\zeta_{\mathrm{RR}}(y)$.

If the final object of interest is $\overline{\zeta}_{\mathrm{RR}}$, i.e. the aggregated version of $\log \theta_{\mathrm{RR}}(x)$ over the entire population, then design 2 is clearly more convenient than design 1. In the case-control sampling design, we need to compute the weighted average of $\zeta_{\mathrm{RR}}(0)$ and $\zeta_{\mathrm{RR}}(1)$. If p^* is known, then conducting inference on $\overline{\zeta}_{\mathrm{RR}}$ is not hard: all of our discussion above applies again, though we need to use the standard error of the linear combination of $\widehat{\beta}_{\text{DML}}(0)$ and $\widehat{\beta}_{\text{DML}}(1)$. More realistically, the only information available to a researcher may be $p^* \in [0, \overline{p}]$ for some known upper bound $\overline{p}$. Then, the sharp bounds for $\overline{\xi}_{\text{RR}}$ will be given by

$$0 \leq \overline{\xi}_{\text{RR}} \leq \max\{\beta(0),\ \beta(0)(1-\overline{p}) + \beta(1)\overline{p}\}. \tag{26}$$

Equation (26) suggests that we can implement "union bounds" to obtain a confidence interval for $\overline{\xi}_{\text{RR}}$. Specifically, we first check if $\beta(0) \geq \beta(1)$ by comparing their estimates. If so, then we use the estimate of $\beta(0)$ and its standard error to compute a one-sided confidence interval. If not, then we use the estimate of $\beta(0)(1-\overline{p}) + \beta(1)\overline{p}$ and its standard error.

6. An Empirical Example

In this section, we provide an empirical example to illustrate the usefulness of our approach. We revisit the ACS 2018 sample extract in Introduction and add covariates to implement the estimation methods we have proposed in this paper. Recall that the sample is restricted to white males residing in California with at least a bachelor's degree. The case sample ($Y = 1$) is composed of 921 individuals whose income is top-coded. To mimic design 1, the control sample ($Y = 0$) of equal size is randomly drawn without replacement from the pool of individuals whose income is not top-coded. Thus, by design, $\mathbb{P}(Y = 1) = 0.5$ and $\hat{h} = 0.5$. Covariates (X) include age and industry codes, and the binary treatment (T) is defined to be one if an individual has a degree beyond bachelor's. Age is restricted to be between 25 and 70.

We consider two different estimators: (i) retrospective sieve logit and (ii) retrospective DML estimator. For (i), only age is included as a covariate with cubic B-splines having three inner knots.[5] For (ii), both age and industry codes are used. In particular, cubic B-splines of age with 17 inner knots (hence, $J_n = 20$) as well as 254 industry dummies are included in this specification, which can be viewed

[5]Specifically, they are 34, 45 and 55, which correspond to 0.25, 0.50 and 0.75 quantiles of the empirical age distribution.

as a high-dimensional setting.[6] Specifically, we implement ℓ_1-penalized logistic estimation with `glmnet` package in R (Friedman, Hastie, and Tibshirani, 2010) to estimate $\mathbb{P}(T = 1|X = x, Y = y), y = 0, 1$ and $\mathbb{P}(Y = 1|X = x)$ with 5-fold cross-fitting. The underlying assumption here is that the B-spline terms plus the industry dummies are rich enough to approximate $\mathbb{P}(T = 1|X = x, Y = y)$ as well as $\mathbb{P}(Y = 1|X = x)$. The penalization tuning parameter is chosen by cross-validation (that is, `lambda.min` in the `glmnet` package). To present a representative result, we draw the control sample 100 times and compute estimates for each draw. Estimates and standard errors reported below are median values out of 100 replications.

TABLE 2. Empirical Results: Sieve Logit

Panel A.	$\beta(1)$	$\beta(0)$
Retrospective Estimate	0.656	0.489
	(0.101)	(0.167)

Note. Standard errors are in the parentheses.

Panel B.	$\exp[\beta(1)]$	$\exp[\beta(0)]$
Retrospective Estimate	1.927	1.631
95% Confidence Interval	[1,2.276]	[1,2.147]

Note. Confidence intervals are obtained under the assumption that the point estimate is the upper bound of $\exp[\beta(y)], y = 0, 1$.

Table 2 reports estimation results with sieve logit estimation. Looking at Panel A, the retrospective sieve estimate of $\beta(1)$ is 0.656, which is larger than that of $\beta(0)$, thereby suggesting that there is heterogeneity among individuals. However, the standard error of $\widehat{\beta}(0)$ is larger than that of $\widehat{\beta}(1)$, which indicates that the difference between the two estimates might be driven by sampling uncertainty. In Panel B, we present point estimates of $\exp[\beta(y)], y = 0, 1$ and their confidence intervals under the assumption that the point estimate is the upper bound because the MTR

[6] Sieve logit estimation without penalization produced bogus results.

and MTS assumptions are more plausible than strong ignorability in this example. The estimates of $\exp[\beta(y)]$ are comparable to the usual odds ratio in terms of its scale; therefore, they can be interpreted similarly. For example, 1.927 of $\exp[\widehat{\beta}(1)]$ roughly means that obtaining a higher-level degree doubles the upper bound for the chance of earning very high incomes. The end point of the confidence interval ranges from 2.15 to 2.28, which includes the unconditional odds ratio of 2.19 using the full sample.

TABLE 3. Empirical Results: Retrospective DML Estimator

Panel A.	$\beta(1)$	$\beta(0)$
Retrospective Estimate	0.816	0.663
	(0.145)	(0.124)

Standard errors are in the parentheses.

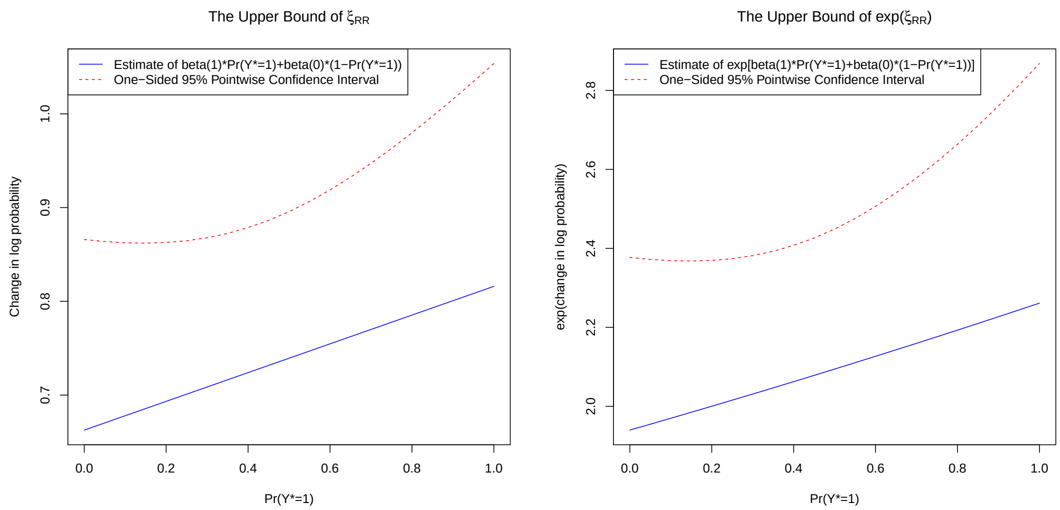
Panel B.	$\exp[\beta(1)]$	$\exp[\beta(0)]$
Retrospective Estimate	2.261	1.940
95% Confidence Interval	[1,2.868]	[1,2.377]

Note. Confidence intervals are obtained under the assumption that the point estimate is the upper bound of $\exp[\beta(y)], y = 0, 1$.

Table 3 reports estimation results with the retrospective DML estimator. The point estimates are larger than those in table 2, indicating that the effect of higher educational attainment might be larger. It is impressive that the standard errors are about the same size as those reported in table 2, given that 254 industry dummies are additionally included with more B-spline terms for age.

In semiparametric estimation with sieve approximation of unknown functions, it is necessary to choose the number J_n of approximating terms. Typically, the optimal choice of J_n for semiparametric estimation is different from one for non-parametric estimation. Furthermore, unlike age, there is no natural ordering in industry codes; thus, it would require an ad hoc grouping of industry dummies to reduce the number of covariates if a researcher needs to use logistic regression

FIGURE 1. The Upper Bounds of $\overline{\xi}_{\text{RR}}$ and $\exp(\overline{\xi}_{\text{RR}})$: Sensitivity Analysis



Note. $\overline{\xi}_{\text{RR}} = \mathbb{E}\big[\log \mathbb{P}\{Y^*(1) = 1|X^*\}\big] - \mathbb{E}\big[\log \mathbb{P}\{Y^*(0) = 1|X^*\}\big]$. The left panel shows the estimate and 95% one-sided pointwise confidence interval for $\overline{\xi}_{\text{RR}}$, as a function of $\mathbb{P}(Y^* = 1)$, and the right panel those for $\exp(\overline{\xi}_{\text{RR}})$.

without penalization. Alternatively, a researcher might want to use machine learning methods to deal with high-dimensionality of B-spline terms and full industry codes. However, it could lead to a question whether and how to conduct inference if one mainly cares about parameters such as $\widehat{\beta}(y)$. The retrospective DML estimation method provides a constructive and affirmative answer to this question.

We end this section by illustrating a sensitivity analysis for $\overline{\xi}_{\text{RR}}$. The left panel of figure 1 shows the estimate and 95% one-sided pointwise confidence interval for $\overline{\xi}_{\text{RR}}$, as a function of $\mathbb{P}(Y^* = 1)$, and the right panel those for $\exp(\overline{\xi}_{\text{RR}})$. In the case-control sampling, the true value of $\mathbb{P}(Y^* = 1)$ may be unknown; however, as we can see from figure 1, we can trace out $\overline{\xi}_{\text{RR}}$ as a function of $\mathbb{P}(Y^* = 1)$, thereby providing a tool for the sensitivity analysis. In the range of $\mathbb{P}(Y^* = 1)$ from 0 to 0.5, the upper end point of the 95% pointwise confidence interval for $\overline{\xi}_{\text{RR}}$ (respectively, $\exp(\overline{\xi}_{\text{RR}})$) is at most 0.9 (respectively, 2.5). Roughly speaking, this implies that it is highly unlikely that obtaining a degree beyond bachelor's improves the chance of earning high incomes by more than a factor of 2.5.

7. Related Literature and Future Research

The literature on causal inference using observational data is vast and the literature on non-random sampling is extensive. In this section we discuss some of the important papers in the context of what we have achieved in this paper.

We have labeled designs 1 and 2 together as Bernoulli sampling, which is the term that we borrowed from Breslow, Robins, and Wellner (2000). The two sampling schemes have been studied under different names by other authors. For instance, Imbens and Lancaster (1996) refer to design 1 as multinomial sampling, and Lancaster and Imbens (1996) call design 2 case-control sampling with contamination, which is borrowed from Heckman and Robb (1985).

The objective of Heckman and Robb (1985) is to estimate the impact of training on earnings under various data scenarios. In that study they discuss common data problems such as oversampling of trainees or "contamination" in the control group, i.e. the training status of the individuals in the control group being unknown. Although the sampling schemes of Heckman and Robb (1985) are similar to designs 1 and 2, they are distinct in the sense that they are not outcome-based but *treatment-based* sampling. In our context, having a control group drawn from the whole population without conditioning on the outcome status makes it easier, not harder, to identify the causal relative risk parameter. For this reason we have referred to design 2 as case-population sampling in order to remove connotations of negativeness from the word "contamination."

Estimating the average treatment effect under treatment-based sampling has been studied by other authors as well. For instance, Heckman and Todd (2009) point out that a matching estimator can be implemented by using the odds ratio of the propensity score fit on the sample because it is a monotone transformation of the true propensity scores. Kennedy, Sjölander, and Small (2015) show that one can estimate the average treatment effect on the treated without the knowledge of the true population probability of the treatment. Assuming the latter is known, Hu and Qin (2018) and Zhang, Hu, and Liu (2019) have developed weighted estimators of the average treatment effect. However, all these methods are based on strong ignorability, and to the best of our knowledge, we are not aware of any work that does not rely on it. We leave it for future research how to extend the approach taken in this paper to the context of treatment-based sampling.

The term Bernoulli sampling has been alternatively used by e.g. Kalbfleisch and Lawless (1988) to describe the case where an individual unit is randomly drawn from the entire population but it is retained or discarded with stratum-specific probabilities. Imbens and Lancaster (1996) use the same terminology, while they call our design 1 multinomial sampling as we mentioned earlier. The case where a given number of observations are randomly drawn from each stratum has been traditionally called the classical stratified sampling scheme (e.g. Hausman and Wise, 1981). However, Imbens and Lancaster (1996) have shown that there is no meaningful difference among the three schemes in that they lead to the same likelihood function to estimate the parameters that appear in the choice probabilities. Since this paper is concerned about a binary outcome, Bernoulli sampling seems more appropriate than multinomial sampling.

In the literature on choice-based sampling, the objective is usually efficiently estimating the parameters that appear in the parametrically specified prospective probabilities. Manski and Lerman (1977) propose a weighted likelihood approach for this purpose under outcome-based sampling. Cosslett (1981) shows that it is feasible to compute the full maximum likelihood estimator. By far the most common specification is the logistic model. However, as Xie and Manski (1989) point out, the logit model can be quite misleading under outcome-based sampling, if the truth is not logistic. Despite its convenience, the logistic specification imposes restrictions on the form of heterogeneity in the causal effect. In contrast, our approach does not restrict the shape of the causal relative risk function $\theta_{\mathrm{RR}}(\cdot)$, thereby allowing an unrestricted form of heterogeneity in the causal treatment effect.

Many papers in this literature use the term "semiparametric" to describe the fact that the marginal distribution of the regressors are left unspecified in their analysis, while the prospective probability, i.e. the conditional distribution of the outcome given the regressors, is still parametric: see e.g. Imbens and Lancaster (1996) and Breslow, Robins, and Wellner (2000). By contrast, our approach is semi-nonparametric in the sense of X. Chen (2007) because we do not impose parametric restrictions anywhere. Instead of relying on the parametric assumption, we directly target the aggregated log odds ratio as the estimand of interest, we articulate its relationship with the fundamental causal parameter of interest, and we have derived the efficiency bound for the estimand under Bernoulli sampling. By combining all these results we can draw robust and efficient inferences on the causal parameter of interest.

In the statistics and epidemiology literature, misspecification and robustness has been addressed from a different perspective. For instance, H.Y. Chen (2007) considers estimating the parameters that appear in the odds ratio in such a way that consistency and asymptotic normality follows as long as either the prospective or the retrospective probability is correctly specified: this approach is known as a doubly robust estimation method. Tchetgen Tchetgen, Robins, and Rotnitzky (2010) take a similar approach, but their estimator is simpler to implement than H.Y. Chen (2007)'s; it is then further operationalized by Tchetgen Tchetgen (2013) under the finite-dimensional logistic assumption. Our estimating equation in (16) is different because our parameter of interest is semi-nonparametric. It is also noteworthy that statisticians and epidemiologists have maintained an active research agenda in case-control studies unlike econometricians. In addition to the aforementioned papers, for instance, Zhou, Herring, Bhattacharya, Olshan, Dunson, and Study (2016) investigate how to deal with high dimensional predictors in the case-control setup using a nonparametric Bayesian approach.

Finally, our causal parameter is defined by a ratio, but it is probably fair to say that a difference (attributable risk in our setup) is a more common measure in econometrics (e.g. Hahn, 1998; Hirano, Imbens, and Ridder, 2003). We do this only because the ratio is mathematically more convenient under outcome-based sampling thanks to the invariance property of the odds ratio. However, it has long been questioned whether the emphasis on relative risk combined with the rare-disease assumption is relevant for public policies: see, e.g., Hsieh, Manski, and McFadden (1985) and Manski (2009) among others. We take a pragmatic approach to this debate and believe that both attributable risk and relative risk are useful for evidence-based policymaking. We plan to work out details for causal attributable risk in a separate paper since its analysis is sufficiently distinct from that of causal relative risk.

Appendix A. Averaging without Taking the Logarithm

In the main text our key estimand was an aggregated version of the *logarithm* of the odds ratio, i.e. $\beta(y) = \mathbb{E}[\log\{\mathrm{OR}(X)\}|Y = y]$ for $y = 0, 1$. As a result, the central causal parameter $\bar{\zeta}_{\mathrm{RR}}$ was defined in (24) by the *logarithm* of relative risk.

Alternatively, one may want to proceed without taking the logarithm in which case we are led to consider

$$\bar{\zeta}_{\mathrm{RR}} := \mathbb{E}\big[\theta_{\mathrm{RR}}(X^*)\big], \;\; \zeta_{\mathrm{RR}}(y) := \int_{\mathcal{X}} \theta_{\mathrm{RR}}(x) dF_{X|Y}(x|y), \;\; \kappa(y) := \int_{\mathcal{X}} \mathrm{OR}(x) dF_{X|Y}(x|y)$$

for $y = 0, 1$. Again, if the MTR and MTS conditions are satisfied, then we have

$$1 \leq \zeta_{\mathrm{RR}}(y) \leq \kappa(y) \tag{27}$$

under both designs 1 and 2, where the inequalities are sharp.

Efficient estimation of $\kappa(y)$ can be explored exactly in the same way as in section 4. Below we present the formula of the efficient influence function, which is an analog of theorem 3.

Theorem A.1. *Suppose that assumptions A, F and G hold and that we have a sample by Bernoulli sampling. Then, for $y = 0, 1$, $\kappa(y)$ is pathwise differentiable and its pathwise* *derivative is given by*

$$K_y(Y,T,X) = \frac{Y^y(1-Y)^{1-y}}{h_0^y(1-h_0)^{1-y}}\Big\{\mathrm{OR}(X) - \kappa(y)\Big\} \\ - \mathrm{OR}(X)\frac{\Delta_0(Y,T,X)}{(1-h_0)w(X)^y} + \mathrm{OR}(X)\frac{w(X)^{1-y}\Delta_1(Y,T,X)}{h_0}.$$

Further, K_y is an element of the tangent space, and therefore, the semiparametric efficiency bound for $\kappa(y)$ is given by $\mathbb{E}\{K_y^2(Y,T,X)\}$.

We omit the proof of theorem A.1 because it is essentially identical to that of Theorem 3. We can construct efficient estimators of $\kappa(y)$ and carry out causal inference on $\overline{\zeta}_{\mathrm{RR}}$ by methods identical to those used in section 4. We do not repeat all the details for brevity.

In general we have the relationship

$$\overline{\xi}_{\mathrm{RR}} \leq \log(\overline{\zeta}_{\mathrm{RR}})$$

by Jensen's inequality. We have chosen $\overline{\xi}_{\mathrm{RR}}$ as our central causal parameter to focus on in the main text because (i) it corresponds to the usual parameter when a parametric logistic regression model is used, and (ii) an average of the log odds ratio is less likely to be affected unduly by outliers than that of the odds ratio itself.

Appendix B. Auxiliary lemmas

Lemma B.1. *Suppose that assumption D holds. Then, for $t = 0, 1$ and for all $x \in \mathcal{X}$,*

$$(-1)^t\big[\mathbb{P}\{Y^*(t) = 1|X^* = x\} - \mathbb{P}(Y^* = 1|X^* = x)\big] \leq 0,$$

where the bounds are sharp.

Proof. Since the two inequalities are similar, we focus on the case of $t = 1$. In this case, the claimed inequality follows from

$$\mathbb{P}\{Y^*(1) = 1, T^* = 1|X^* = x\} + \mathbb{P}\{Y^*(1) = 1, T^* = 0|X^* = x\} \\ \geq \mathbb{P}\{Y^*(1) = 1, T^* = 1|X^* = x\} + \mathbb{P}\{Y^*(0) = 1, T^* = 0|X^* = x\}.$$

For sharpness, we know from assumption D that

$$\mathbb{P}\{Y^*(1) = 1, T^* = 0|X^* = x\} - \mathbb{P}\{Y^*(0) = 1, T^* = 0|X^* = x\} \\ = \mathbb{P}\{Y^*(1) = 1, Y^*(0) = 0, T^* = 0|X^* = x\},$$

where the right–hand side is unrestricted between 0 and 1. □

Lemma B.2. *Suppose that assumption E holds. Then, for $t = 0, 1$ and for all $x \in \mathcal{X}$,*

$$(-1)^t \left[\mathbb{P}\{Y^*(t) = 1|X^* = x\} - \mathbb{P}(Y^* = 1|T^* = t, X^* = x)\right] \geq 0,$$

where the bounds are sharp. Furthermore, if $0 < \mathbb{P}(T^* = 1|X^* = x) < 1$, these inequalities hold with equality if and only if assumption E is satisfied with equality.

Proof. Since the two inequalities are similar, we focus on the case of $t = 1$. First,

$$\mathbb{P}\{Y^*(1) = 1|X^* = x\} = \mathbb{P}(Y^* = 1|T^* = 1, X^* = x)\mathbb{P}(T^* = 1|X^* = x) \\ + \mathbb{P}\{Y^*(1) = 1|T^* = 0, X^* = x\}\mathbb{P}(T^* = 0|X^* = x), \quad (28)$$

where we note from assumption E that there exists some $C_x \in [0, 1]$ such that

$$\mathbb{P}(Y^* = 1|T^* = 1, X^* = x) = \mathbb{P}\{Y^*(1) = 1|T^* = 0, X^* = x\} + C_x. \quad (29)$$

Combining equations (28) and (29) yields the first inequality in the lemma statement. Therefore,

$$\mathbb{P}\{Y^*(1) = 1|X^* = x\} = \mathbb{P}(Y^* = 1|T^* = 1, X^* = x) - C_x \cdot \mathbb{P}(T^* = 0|X^* = x) \\ \leq \mathbb{P}(Y^* = 1|T^* = 1, X^* = x). \quad (30)$$

Sharpness follows from the fact that C_x is not restricted except that it is between 0 and 1. Also, if $\mathbb{P}(T^* = 0|X^* = x) > 0$, then the last inequality in equation (30) holds with equality if and only if $C_x = 0$. □

APPENDIX C. PROOFS OF THE RESULTS IN THE MAIN TEXT

Proof of Lemma 1: By the Bayes rule,

$$\text{OR}(x) = \frac{\mathbb{P}(T = 1|X = x, Y = 1)}{\mathbb{P}(T = 0|X = x, Y = 1)} \frac{\mathbb{P}(T = 0|X = x, Y = 0)}{\mathbb{P}(T = 1|X = x, Y = 0)}.$$

Then, under design 1, for all $x \in \mathcal{X}$,

$$\text{OR}(x) = \frac{\mathbb{P}(T^* = 1|X^* = x, Y^* = 1)}{\mathbb{P}(T^* = 0|X^* = x, Y^* = 1)} \frac{\mathbb{P}(T^* = 0|X^* = x, Y^* = 0)}{\mathbb{P}(T^* = 1|X^* = x, Y^* = 0)} = \text{OR}^*(x),$$

where the second equality again follows from the Bayes rule. Now, under design 2, for all $x \in \mathcal{X}$,

$$\text{OR}(x) = \frac{\mathbb{P}(T^* = 1|X^* = x, Y^* = 1)}{\mathbb{P}(T^* = 0|X^* = x, Y^* = 1)} \frac{\mathbb{P}(T^* = 0|X^* = x)}{\mathbb{P}(T^* = 1|X^* = x)} = \text{RR}^*(x). \qquad \square$$

Proof of Lemma 2: Let γ be the parameter denoting regular parametric submodels, where the true value will be denoted by γ_0. Then, by using the likelihood function in equation (11), the score evaluated at γ_0 is equal to

$$\begin{aligned} &(1 - Y)\Big[S_{X|Y}(X|0) + \frac{\{T - \mathbb{P}(T = 1|X, Y = 0)\}\partial_\gamma \mathbb{P}(T = 1|X, Y = 0; \gamma_0)}{\mathbb{P}(T = 1|X, Y = 0)\{1 - \mathbb{P}(T = 1|X, Y = 0)\}}\Big] \\ &\quad + Y\Big[S_{X|Y}(X|1) + \frac{\{T - \mathbb{P}(T = 1|X, Y = 1)\}\partial_\gamma \mathbb{P}(T = 1|X, Y = 1; \gamma_0)}{\mathbb{P}(T = 1|X, Y = 1)\{1 - \mathbb{P}(T = 1|X, Y = 1)\}}\Big], \end{aligned} \tag{31}$$

where $S_{X|Y}(x|y) = \partial_\gamma \log f_{X|Y}(x|y; \gamma_0)$ is restricted only by $\mathbb{E}\{S_{X|Y}(X|y)|Y = y\} = 0$, while the derivatives $\partial_\gamma \mathbb{P}(T = 1|X, Y = y, \gamma_0)$ are unrestricted. □

Proof of Theorem 1: In view of Lemma 1, the theorem follows immediately since

$$\mathbb{P}(Y^* = 1|T^* = t, X^* = x) = \mathbb{P}\{Y^*(t) = 1|T^* = t, X^* = x\} = \mathbb{P}\{Y^*(t) = 1|X^* = x\},$$

where the last equality is by the assumption of unconfoundedness. □

Proof of Theorem 2: *Part (i).* The sharp lower bound of $\theta_{\text{RR}}(x)$ follows from lemma B.1. To prove that $\theta_{\text{RR}}(x) \leq \theta_{\text{OR}}(x)$ for all $x \in \mathcal{X}$, note that

$$\frac{\theta_{\text{OR}}(x)}{\theta_{\text{RR}}(x)} = \frac{\mathbb{P}\{Y^*(0) = 0|X^* = x\}}{\mathbb{P}\{Y^*(1) = 0|X^* = x\}} \geq 1$$

since by lemma B.1,

$$(-1)^t \left[\mathbb{P}\{Y^*(t) = 0|X^* = x\} - \mathbb{P}(Y^* = 0|X^* = x)\right] \geq 0 \quad \text{for } t = 0, 1.$$

Part (ii). The sharp upper bound of $\theta_{\text{RR}}(x)$ under design 2 follows from lemmas 1 and B.2 because

$$\theta_{\text{RR}}(x) \leq \frac{\mathbb{P}(Y^* = 1|T^* = 1, X^* = x)}{\mathbb{P}(Y^* = 1|T^* = 0, X^* = x)} = \text{RR}^*(x) = \text{OR}(x). \tag{32}$$

The case of $\theta_{\text{OR}}(x)$ under design 1 similarly uses the fact that lemma B.2 yields

$$\frac{\mathbb{P}\{Y^*(0) = 0|X^* = x\}}{\mathbb{P}\{Y^*(1) = 0|X^* = x\}} \leq \frac{\mathbb{P}(Y^* = 0|T^* = 0, X^* = x)}{\mathbb{P}(Y^* = 0|T^* = 1, X^* = x)}. \tag{33}$$

Combining equation (33) with (32) yields that under design 1,

$$\theta_{\text{OR}}(x) \leq \text{OR}^*(x) = \text{OR}(x).$$

Part (iii). The final statement follows immediately from lemma B.2. □

Proof of theorem 3: For brevity, we focus on $\beta(0)$ and let $\beta = \beta(0)$. Proof for $\beta(1)$ is analogous. Let $p_0(x) = \mathbb{P}(T = 1|X = x, Y = 0)$ and $p_1(x) = \mathbb{P}(T = 1|X = x, Y = 1)$. Note that

$$\beta(\gamma) = \int_{\mathcal{X}} \log \Bigg[\underbrace{\frac{p_1(x;\gamma)}{1 - p_1(x;\gamma)} \cdot \frac{1 - p_0(x;\gamma)}{p_0(x;\gamma)}}_{:=\text{OR}(x;\gamma)} \Bigg] f_{X|Y}(x|0;\gamma)dx, \tag{34}$$

where γ represents regular parametric submodels such that γ_0 is the truth. Then,

$$\begin{aligned}\partial_\gamma \mathrm{OR}(x;\gamma_0) &= \partial_\gamma p_1(x;\gamma_0)\frac{\{1-p_0(x)\}}{p_0(x)\{1-p_1(x)\}^2} - \partial_\gamma p_0(x;\gamma_0)\frac{p_1(x)}{p_0^2(x)\{1-p_1(x)\}} \\ &= \frac{\partial_\gamma p_1(x;\gamma_0)}{p_1(x)\{1-p_1(x)\}}\mathrm{OR}(x) - \frac{\partial_\gamma p_0(x;\gamma_0)}{p_0(x)\{1-p_0(x)\}}\mathrm{OR}(x). \quad (35)\end{aligned}$$

Therefore,

$$\begin{aligned}\partial_\gamma \beta(\gamma_0) &= \int \Big[\frac{\partial_\gamma \mathrm{OR}(x;\gamma_0)}{\mathrm{OR}(x)} + \log\{\mathrm{OR}(x)\}S_{X|Y}(x|0)\Big] f_{X|Y}(x|0)dx \\ &= \int \Big[\frac{\partial_\gamma p_1(x;\gamma_0)}{p_1(x)\{1-p_1(x)\}} - \frac{\partial_\gamma p_0(x;\gamma_0)}{p_0(x)\{1-p_0(x)\}} + \log\{\mathrm{OR}(x)\}S_{X|Y}(x|0)\Big] f_{X|Y}(x|0)dx.\end{aligned} \quad (36)$$

Now, we only need to verify the equality between $\mathbb{E}\{F_0(Y,T,X)S(Y,T,X)\}$ and

$$\int \Big[\underbrace{\frac{\partial_\gamma p_1(x;\gamma_0)}{p_1(x)\{1-p_1(x)\}}}_{:=A_1(x)} - \underbrace{\frac{\partial_\gamma p_0(x;\gamma_0)}{p_0(x)\{1-p_0(x)\}}}_{:=A_0(x)} + \log\{\mathrm{OR}(x)\}S_{X|Y}(x|0)\Big] f_{X|Y}(x|0)dx, \quad (37)$$

where $F_0(Y,T,X)$ and $S(Y,T,X)$ are given in the theorem statement and equation (31), respectively: i.e.

$$\begin{aligned}&S(Y,T,X) \\ &= (1-Y)\Big[S_{X|Y}(X|0) + \big\{T-p_0(X)\big\}A_0(X)\Big] + Y\Big[S_{X|Y}(X|1) + \big\{T-p_1(X)\big\}A_1(X)\Big], \\ &F_0(Y,T,X) \\ &= \frac{1-Y}{1-h_0}\left[\log \mathrm{OR}(X) - \beta - \frac{\{T-p_0(X)\}}{p_0(X)\{1-p_0(X)\}}\right] + \frac{Y}{h_0}\frac{f_{X|Y}(X|0)}{f_{X|Y}(X|1)}\frac{\{T-p_1(X)\}}{p_1(X)\{1-p_1(X)\}}.\end{aligned}$$

Note that $F_0(Y, T, X)S(Y, T, X)$ is equal to

$$\frac{1-Y}{1-h_0}\left[\log \mathrm{OR}(X) - \beta - \frac{\{T - p_0(X)\}}{p_0(X)\{1-p_0(X)\}}\right]\left[S_{X|Y}(X|0) + \{T - p_0(X)\}A_0(X)\right]$$
$$+ \frac{Y}{h_0}\frac{f_{X|Y}(X|0)}{f_{X|Y}(X|1)}\left[\frac{\{T - p_1(X)\}}{p_1(X)\{1-p_1(X)\}}\right]\left[S_{X|Y}(X|1) + \{T - p_1(X)\}A_1(X)\right].$$

Here, taking expectations directly shows that $\mathbb{E}\{F_0(Y, T, X)S(Y, T, X)\}$ is equal to

$$\mathbb{E}\{\log\{\mathrm{OR}(X)\}S_{X|Y}(X|0) - A_0(X)|Y=0\} + \mathbb{E}\left\{\frac{f_{X|Y}(X|0)}{f_{X|Y}(X|1)}A_1(X)\middle|Y=1\right\},$$

which is equal to the expression in equation (37) since

$$\mathbb{E}\left\{\frac{f_{X|Y}(X|0)}{f_{X|Y}(X|1)}A_1(X)\middle|Y=1\right\} = \mathbb{E}\{A_1(X)|Y=0\}.$$

Finally, it follows from lemma 2 that F_0 is an element of the tangent space. □

Proofs of theorems 4 and 5 are provided in appendix S-1, which is only for online: The proof of theorem 4 is similar to that of theorem 3. The proof of theorem 5 does not provide any additional insight above DML. □

References

ACKERBERG, D., X. CHEN, AND J. HAHN (2012): "A practical asymptotic variance estimator for two-step semiparametric estimators," *Review of Economics and Statistics*, 94(2), 481–498.

ACKERBERG, D., X. CHEN, J. HAHN, AND Z. LIAO (2014): "Asymptotic efficiency of semiparametric two-step GMM," *Review of Economic Studies*, 81(3), 919–943.

AI, C., AND X. CHEN (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.

——— (2012): "The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions," *Journal of Econometrics*, 170(2), 442–457.

BELLONI, A., V. CHERNOZHUKOV, AND Y. WEI (2016): "Post-Selection Inference for Generalized Linear Models With Many Controls," *Journal of Business & Economic Statistics*, 34(4), 606–619.

BHATTACHARYA, J., A. M. SHAIKH, AND E. VYTLACIL (2008): "Treatment effect bounds under monotonicity assumptions: an application to Swan-Ganz catheterization," *American Economic Review: Papers and Proceedings*, 98(2), 351–56.

BHATTACHARYA, J., A. M. SHAIKH, AND E. VYTLACIL (2012): "Treatment effect bounds: An application to Swan-Ganz catheterization," *Journal of Econometrics*, 168(2), 223–243.

BRESLOW, N. E. (1996): "Statistics in epidemiology: the case-control study," *Journal of the American Statistical Association*, 91(433), 14–28.

BRESLOW, N. E., AND N. E. DAY (1980): *Statistical Methods in Cancer Research I. The Analysis of Case-Control Studies*, vol. 1. International Agency for Research on Cancer, Lyon, France.

BRESLOW, N. E., J. M. ROBINS, AND J. A. WELLNER (2000): "On the semiparametric efficiency of logistic regression under case-control sampling," *Bernoulli*, 6(3), 447–455.

CARVALHO, L. S., AND R. R. SOARES (2016): "Living on the edge: Youth entry, career and exit in drug-selling gangs," *Journal of Economic Behavior & Organization*, 121, 77–98.

CHEN, K. (2001): "Parametric models for response-biased sampling," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4), 775–789.

CHEN, X. (2007): "Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models," vol. 6 of *Handbook of Econometrics*, pp. 5549–5632. Elsevier.

CHEN, H.Y. (2007): "A semiparametric odds ratio model for measuring association," *Biometrics*, 63(2), 413–421.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DIMIRER, E. DUFLO, C. HANSEN, W. NEWEY, AND J. ROBINS (2018): "Double/debiased machine learning for treatment and structural parameters," *Econometrics Journal*, 21, C1–C68.

CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): "Intersection bounds: Estimation and inference," *Econometrica*, 81(2), 667–737.

CORNFIELD, J. (1951): "A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix," *Journal of the National Cancer Institute*, 11(6), 1269–1275.

COSSLETT, S. R. (1981): "Maximum Likelihood Estimator for Choice-Based Samples," *Econometrica*, 49(5), 1289–1316.

CURRIE, J., AND M. NEIDELL (2005): "Air pollution and infant health: what can we learn from California's recent experience?," *Quarterly Journal of Economics*, 120(3), 1003–1030.

DOMOWITZ, I., AND R. L. SARTAIN (1999): "Determinants of the Consumer Bankruptcy Decision," *Journal of Finance*, 54(1), 403–420.

FRIEDMAN, J., T. HASTIE, AND R. TIBSHIRANI (2010): "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33(1), 1–22.

HAHN, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

HAUSMAN, J. A., AND D. A. WISE (1981): "Stratification on endogenous variables and estimation: The Gary income maintenance experiment," *Structural analysis of discrete data with econometric applications*, pp. 365–391.

HECKMAN, J. J., AND R. ROBB (1985): "Alternative methods for evaluating the impact of interventions: An overview," *Journal of econometrics*, 30(1-2), 239–267.

HECKMAN, J. J., AND P. E. TODD (2009): "A note on adapting propensity score matching and selection models to choice based samples," *Econometrics Journal*, 12(s1), S230–S234.

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

HOLLAND, P. W., AND D. B. RUBIN (1988): "Causal Inference in Retrospective Studies," *Evaluation Review*, 12(3), 203–231.

HSIEH, D. A., C. F. MANSKI, AND D. MCFADDEN (1985): "Estimation of Response Probabilities from Augmented Retrospective Observations," *Journal of the American Statistical Association*, 80(391), 651–662.

HU, Z., AND J. QIN (2018): "Generalizability of causal inference in observational studies under retrospective convenience sampling," *Statistics in Medicine*, 37(19), 2874–2883.

IMBENS, G. W. (1992): "An Efficient Method of Moments Estimator for Discrete Choice Models With Choice-Based Sampling," *Econometrica*, 60(5), 1187–1214.

IMBENS, G. W., AND T. LANCASTER (1996): "Efficient estimation and stratified sampling," *Journal of Econometrics*, 74(2), 289–318.

IMBENS, G. W., AND D. B. RUBIN (2015): *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

JUN, S. J., AND S. LEE (2019): "Identifying the effect of persuasion," arXiv:1812.02276.

KALBFLEISCH, J., AND J. LAWLESS (1988): "Estimation of Reliability in Field–Performance Studies," *Technometrics*, 30(4), 365–378.

KENNEDY, E. H., A. SJÖLANDER, AND D. SMALL (2015): "Semiparametric causal inference in matched cohort studies," *Biometrika*, 102(3), 739–746.

KIM, W., K. KWON, S. KWON, AND S. LEE (2018): "The identification power of smoothness assumptions in models with counterfactual outcomes," *Quantitative Economics*, 9(2), 617–642.

KREIDER, B., J. V. PEPPER, C. GUNDERSEN, AND D. JOLLIFFE (2012): "Identifying the Effects of SNAP (Food Stamps) on Child Health Outcomes When Participation Is Endogenous and Misreported," *Journal of the American Statistical Association*, 107(499), 958–975.

LANCASTER, T., AND G. IMBENS (1996): "Case-control studies with contaminated controls," *Journal of Econometrics*, 71(1-2), 145–160.

LEE, S., K. SONG, AND Y.-J. WHANG (2018): "Testing for a general class of functional inequalities," *Econometric Theory*, 34(5), 1018–1064.

MACHADO, C., A. M. SHAIKH, AND E. J. VYTLACIL (2019): "Instrumental variables and the sign of the average treatment effect," *Journal of Econometrics*, 212(2), 522–555.

MANSKI, C. F. (1997): "Monotone Treatment Response," *Econometrica*, 65(6), 1311–1334.

MANSKI, C. F. (2003): *Partial identification of probability distributions*. Springer-Verlag.

——— (2009): *Identification for prediction and decision*. Harvard University Press.

MANSKI, C. F., AND S. R. LERMAN (1977): "The Estimation of Choice Probabilities from Choice Based Samples," *Econometrica*, 45(8), 1977–1988.

MANSKI, C. F., AND D. MCFADDEN (1981): "Alternative estimators and sample designs for discrete choice analysis," in *Structural analysis of discrete data with econometric applications*, ed. by C. F. Manski, and D. McFadden, vol. 2, pp. 51–111. MIT Press, Cambridge, MA.

MANSKI, C. F., AND J. V. PEPPER (2000): "Monotone instrumental variables: With an application to the returns to schooling," *Econometrica*, 68(4), 997–1010.

MCFADDEN, D. (2015): "Observational Studies: Outcome-Based Sampling," in *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, ed. by J. D. Wright, pp. 103–106. Elsevier, Oxford.

NEWEY, W. K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

——— (1994): "The asymptotic variance of semiparametric estimators," *Econometrica*, 62(6), 1349–1382.

OKUMURA, T., AND E. USUI (2014): "Concave-monotone treatment response and monotone treatment selection: With an application to the returns to schooling," *Quantitative Economics*, 5(1), 175–194.

RUGGLES, S., S. FLOOD, R. GOEKEN, J. GROVER, E. MEYER, J. PACAS, AND M. SOBEK (2019): "IPUMS USA: Version 9.0 [dataset]," https://doi.org/10.18128/D010.V9.0.

TAMER, E. (2010): "Partial identification in econometrics," *Annual Review of Economics*, 2(1), 167–195.

TCHETGEN TCHETGEN, E. J. (2013): "On a closed-form doubly robust estimator of the adjusted odds ratio for a binary exposure," *American journal of epidemiology*, 177(11), 1314–1316.

TCHETGEN TCHETGEN, E. J., J. M. ROBINS, AND A. ROTNITZKY (2010): "On doubly robust estimation in a semiparametric odds ratio model," *Biometrika*, 97(1), 171–180.

VAN DE GEER, S. A. (2008): "High-dimensional generalized linear models and the lasso," *Annals of Statistics*, 36(2), 614–645.

XIE, J., Y. LIN, X. YAN, AND N. TANG (2019): "Category-Adaptive Variable Screening for Ultra-High Dimensional Heterogeneous Categorical Data," *Journal of the American Statistical Association*, forthcoming.

XIE, Y., AND C. F. MANSKI (1989): "The logit model and response-based samples," *Sociological Methods & Research*, 17(3), 283–302.

ZHANG, Z., Z. HU, AND C. LIU (2019): "Estimating the Population Average Treatment Effect in Observational Studies with Choice-Based Sampling," *The international journal of biostatistics*, 15(1).

ZHOU, J., A. H. HERRING, A. BHATTACHARYA, A. F. OLSHAN, D. B. DUNSON, AND N. B. D. P. STUDY (2016): "Nonparametric Bayes modeling for case control studies with many predictors," *Biometrics*, 72(1), 184–192.

Online Supplement to "Causal Inference in Case-Control Studies" by Jun and Lee

Appendix S-1. Additional Proofs

Proof of theorem 4: For simplicity, in the proof, we focus on $\beta(0)$ and let it be denoted by β. The case of $\beta(1)$ is similar. Recall that

$$\tilde{F}_0(\eta)[Y,T,X]$$
$$= \frac{1-Y}{1-h_0}\left[\log \widetilde{\mathrm{OR}}(\eta_1)[X] - \beta - \frac{\{T-a(X)\}}{a(X)\{1-a(X)\}}\right] + \frac{Y\tilde{w}(\eta_2)[X]}{h_0}\frac{\{T-b(X)\}}{b(X)\{1-b(X)\}}.$$

Note that

$$\mathbb{E}\Big\{\tilde{F}_0(\eta)[Y,T,X]\Big\} = \mathbb{E}\{\log \widetilde{\mathrm{OR}}(\eta_1)[X] - \beta \mid Y=0\}$$
$$+\mathbb{E}\Big\{\tilde{\Delta}_0(\eta)[T,X]\Big| Y=0\Big\} + \mathbb{E}\Big\{\tilde{\Delta}_1(\eta)[T,X]\Big| Y=1\Big\}, \quad \text{(S.1)}$$

where

$$\tilde{\Delta}_0(\eta)[T,X] = -\frac{\{T-a(X)\}}{a(X)\{1-a(X)\}}, \quad \text{(S.2)}$$
$$\tilde{\Delta}_1(\eta)[T,X] = \frac{\tilde{w}(\eta_2)[X]\{T-b(X)\}}{b(X)\{1-b(X)\}}. \quad \text{(S.3)}$$

Here,

$$\mathbb{E}\Big(\partial_\gamma \tilde{\Delta}_0\{\eta_0+\gamma(\eta-\eta_0)\}[T,X]\big|_{\gamma=0} \;\Big|\; X, Y=0\Big) = \frac{a(X)-p_0(X)}{p_0(X)\{1-p_0(X)\}},$$
$$\mathbb{E}\Big(\partial_\gamma \tilde{\Delta}_1\{\eta_0+\gamma(\eta-\eta_0)\}[T,X]\big|_{\gamma=0} \;\Big|\; X, Y=1\Big) = -w(X)\frac{b(X)-p_1(X)}{p_1(X)\{1-p_1(X)\}},$$

where $p_0(X) = \mathbb{P}(T = 1|X, Y = 0)$ and $p_1 = \mathbb{P}(T = 1|X, Y = 1)$ as before. Therefore,

$$\mathbb{E}\Big(\partial_\gamma \tilde{\Delta}_0\{\eta_0 + \gamma(\eta - \eta_0)\}[T, X]\big|_{\gamma=0} \,\Big| Y = 0\Big) = \mathbb{E}\Big\{\frac{a(X) - p_0(X)}{p_0(X)\{1 - p_0(X)\}}\Big| Y = 0\Big\}, \quad \text{(S.4)}$$

and

$$\mathbb{E}\Big(\partial_\gamma \tilde{\Delta}_1\{\eta_0 + \gamma(\eta - \eta_0)\}[T, X]\big|_{\gamma=0} \,\Big|\, Y = 1\Big) = -\mathbb{E}\Big\{w(X)\frac{b(X) - p_1(X)}{p_1(X)\{1 - p_1(X)\}}\Big| Y = 1\Big\} = -\mathbb{E}\Big\{\frac{b(X) - p_1(X)}{p_1(X)\{1 - p_1(X)\}}\Big| Y = 0\Big\}. \quad \text{(S.5)}$$

Now, similarly to equation (35), we have

$$\partial_\gamma \log \widetilde{\mathrm{OR}}\{\eta_{10} + \gamma(\eta_1 - \eta_{10})\}\big|_{\gamma=0} = \frac{b(X) - p_1(X)}{p_1(x)\{1 - p_1(x)\}} - \frac{a(X) - p_0(X)}{p_0(x)\{1 - p_0(x)\}}. \quad \text{(S.6)}$$

Therefore, the conclusion follows from equations (S.1) and (S.4) to (S.6). □

Proof of theorem 5: As in the previous proofs, we focus on $\beta(0) \equiv \beta$. The case of $\beta(1)$ is similar. We verify Assumptions 3.1 and 3.2 of DML (Chernozhukov, Chetverikov, Dimirer, Duflo, Hansen, Newey, and Robins, 2018). Using the notation used in DML,

$$\psi(W; \beta, \eta) = \tilde{F}_y(\eta)[Y, T, X]$$

with $W = (Y, T, X)$. Then our case belongs to that of linear scores, namely

$$\psi(W; \theta, \eta) = \psi^a(W; \eta)\theta + \psi^b(W; \eta),$$

where

$$\psi^a(W;\eta) = -\frac{1-Y}{1-h_0},$$

$$\psi^b(W;\eta) = \frac{1-Y}{1-h_0}\left[\log\widetilde{\mathrm{OR}}(\eta_1)[X] - \frac{T-a(X)}{a(X)\{1-a(X)\}}\right] + \frac{Y\tilde{w}(\eta_2)[X]}{h_0}\frac{T-b(X)}{b(X)\{1-b(X)\}}.$$

Verification of Assumption 3.1 of DML. Under assumptions F and G, Assumption 3.1 of DML is satisfied with $\lambda_N = 0$ and $J_0 = 1$ (using DML's notation). Specifically, Assumption 3.1 (a) of DML is satisfied by (16); part (b) is by the linearity of the score ψ; part (c) is by assumptions F and G; part (d) is by theorem 4; part (e) follows because $\mathbb{E}[\psi^a(W;\eta_0)] = 1$.

Verification of Assumption 3.2 (b) of DML. It holds trivially that $|\psi^a(W;\eta)|$ is bounded by a constant uniformly in η. Moreover, by assumption F, there is a constant $c_1 < \infty$ such that

$$|\psi(W;\beta,\eta)| \leq c_1$$

uniformly in η almost surely.

Verification of Assumption 3.2 (d) of DML. Note that

$$\mathbb{E}[\psi^2(W;\beta,\eta_0)] \geq \frac{1}{1-h_0}\mathbb{E}\left[\{\log\mathrm{OR}(X)-\beta\}^2 + \frac{1}{p_0(X)\{1-p_0(X)\}}\right],$$

which is bounded from below by a constant under assumption F.

Since Assumption 3.2 (a) of DML is the definition of the first stage estimator, theorem 5 follows immediately from Theorems 3.1 and 3.2 of DML, provided that we verify the remaining Assumption 3.2 (c) of DML.

Verification of Assumption 3.2 (c) of DML. Using the notation used in DML, define

$$r_n := \sup_{\eta \in \mathcal{T}_N} |\mathbb{E}[\psi^a(W;\eta) - \psi^a(W;\eta_0)]|,$$

$$r_n' := \sup_{\eta \in \mathcal{T}_N} (\mathbb{E}[|\psi(W;\beta,\eta) - \psi(W;\beta,\eta_0)|^2])^{1/2},$$

$$\lambda_n' := \sup_{\gamma \in (0,1), \eta \in \mathcal{T}_N} |\partial_\gamma^2 \mathbb{E}[\psi(W;\beta,\eta_0 + \gamma(\eta - \eta_0))]|.$$

Step 1. Note that $r_n = 0$ since $\psi^a(W;\eta)$ does not depend on η.

Step 2. Now write that

$$\begin{aligned} \mathbb{E}[|\psi(W;\beta,\eta) - \psi(W;\beta,\eta_0)|^2])^{1/2} &= \|\psi(W;\beta,\eta) - \psi(W;\beta,\eta_0)\|_{P,2} \\ &\leq \|\mathcal{T}_1\|_{P,2} + \|\mathcal{T}_2\|_{P,2}, \end{aligned}$$

where

$$\begin{aligned} \mathcal{T}_1 &:= \frac{1-Y}{1-h_0}\left[\log \widetilde{\mathrm{OR}}(\eta_1)[X] - \frac{T - a(X)}{a(X)\{1-a(X)\}}\right] \\ &\quad - \frac{1-Y}{1-h_0}\left[\log \mathrm{OR}(X) - \frac{T - p_0(X)}{p_0(X)\{1-p_0(X)\}}\right], \\ \mathcal{T}_2 &:= \frac{Y\tilde{w}(\eta_2)[X]}{h_0}\left[\frac{T - b(X)}{b(X)\{1-b(X)\}}\right] - \frac{Yw(X)}{h_0}\left[\frac{T - p_1(X)}{p_1(X)\{1-p_1(X)\}}\right]. \end{aligned}$$

Then, in view of assumptions F and H, there exists a sequence $\tilde{\delta}_n \to 0$ such that

$$\left[\mathbb{E}\{|\psi(W;\beta,\eta) - \psi(W;\beta,\eta_0)|^2\}\right]^{1/2} \leq \tilde{\delta}_n$$

holds with probability at least $1 - \tau_n$. This implies that we can take $r_n' = \tilde{\delta}_n$.

Step 3. Define $a_\gamma(X) := p_0(X) + \gamma\{a(X) - p_0(X)\}$ and $b_\gamma(X) := p_1(X) + \gamma\{b(X) - p_1(X)\}$. Note that

$$\begin{aligned} \partial_\gamma \log \widetilde{\mathrm{OR}}\{\eta_0 + \gamma(\eta - \eta_0)\}[X] &= \frac{\partial_\gamma \widetilde{\mathrm{OR}}\{\eta_0 + \gamma(\eta - \eta_0)\}[X]}{\widetilde{\mathrm{OR}}\{\eta_0 + \gamma(\eta - \eta_0)\}[X]} \\ &= \frac{b(X) - p_1(X)}{b_\gamma(X)\{1 - b_\gamma(X)\}} - \frac{a(X) - p_0(X)}{a_\gamma(X)\{1 - a_\gamma(X)\}}. \end{aligned}$$

In addition,

$$
\begin{aligned}
&\partial_\gamma \left[\frac{\{T - a_\gamma(X)\}}{a_\gamma(X)\{1 - a_\gamma(X)\}} \right] \\
&= -\frac{a(X) - p_0(X)}{a_\gamma(X)\{1 - a_\gamma(X)\}} - \frac{\{T - a_\gamma(X)\}\{1 - 2a_\gamma(X)\}}{a_\gamma^2(X)\{1 - a_\gamma(X)\}^2}\{a(X) - p_0(X)\}, \\
&\partial_\gamma \left[\tilde{w}\{\eta_2 + \gamma(\eta_2 - \eta_{20})\}[X] \frac{T - b_\gamma(X)}{b_\gamma(X)\{1 - b_\gamma(X)\}} \right] \\
&= \frac{T - b_\gamma(X)}{b_\gamma(X)\{1 - b_\gamma(X)\}}(\eta_2 - \eta_{20})[X] - \tilde{w}\{\eta_2 + \gamma(\eta_2 - \eta_{20})\}[X] \frac{b(X) - p_1(X)}{b_\gamma(X)\{1 - b_\gamma(X)\}} \\
&- \tilde{w}\{\eta_2 + \gamma(\eta_2 - \eta_{20})\}[X] \frac{\{T - b_\gamma(X)\}\{1 - 2b_\gamma(X)\}}{b_\gamma^2(X)\{1 - b_\gamma(X)\}^2}\{b(X) - p_1(X)\}.
\end{aligned}
$$

Combining these yields

$$
\begin{aligned}
&\partial_\gamma \psi(W; \beta, \eta_0 + \gamma(\eta - \eta_0)) \\
&= \frac{1 - Y}{1 - h_0} \left[\frac{b(X) - p_1(X)}{b_\gamma(X)\{1 - b_\gamma(X)\}} + \frac{\{T - a_\gamma(X)\}\{1 - 2a_\gamma(X)\}}{a_\gamma^2(X)\{1 - a_\gamma(X)\}^2}\{a(X) - p_0(X)\} \right] \\
&+ \frac{Y}{h_0} \left[\frac{T - b_\gamma(X)}{b_\gamma(X)\{1 - b_\gamma(X)\}}(\eta_2 - \eta_{20})[X] - \tilde{w}\{\eta_2 + \gamma(\eta_2 - \eta_{20})\}[X] \frac{b(X) - p_1(X)}{b_\gamma(X)\{1 - b_\gamma(X)\}} \right. \\
&\quad \left. - \tilde{w}\{\eta_2 + \gamma(\eta_2 - \eta_{20})\}[X] \frac{\{T - b_\gamma(X)\}\{1 - 2b_\gamma(X)\}}{b_\gamma^2(X)\{1 - b_\gamma(X)\}^2}\{b(X) - p_1(X)\} \right].
\end{aligned}
$$

If we take the second-order derivative in the equation above, we can see that each term of the second-order derivatives on the right-hand side can be bounded in absolute value by a constant times $\chi(a, b)$, which is defined to be equal to

$$
\max \left[\{a(X) - p_0(X)\}^2, \{b(X) - p_1(X)\}^2, \{\eta_2(X) - \eta_{20}(X)\}\{b(X) - p_1(X)\} \right].
$$

Therefore, there exists a universal constant $C < \infty$ such that

$$
|\partial_\gamma^2 \mathbb{E}[\psi(W; \beta, \eta_0 + \gamma(\eta - \eta_0))]| \leq C\,\chi(a, b).
$$

Then, by assumption H, there exists a sequence $\tilde{\delta}'_n \to 0$ such that

$$\sup_{\gamma \in (0,1), \eta \in \mathcal{T}_N} |\partial_\gamma^2 \mathbb{E}[\psi(W; \beta, \eta_0 + \gamma(\eta - \eta_0))]| \leq \tilde{\delta}'_n n^{-1/2}$$

holds with probability at least $1 - \tau_n$. Therefore, we can take $\lambda'_n = \tilde{\delta}'_n n^{-1/2}$. □

Reference

Chernozhukov V., D. Chetverikov, M. Dimirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Doubld/debiased machine learning for treatment and structural parameters," *Econometrics Journal*, 21, C1–C68.